A quick, somewhat easy-to-read introduction to empirical social science research methods

A QUICK, SOMEWHAT EASY-TO-READ INTRODUCTION TO EMPIRICAL SOCIAL SCIENCE RESEARCH METHODS

CHRISTOPHER HORNE

University of Tennessee at Chattanooga
Chattanooga (Tenn.)

If you use this text in any way, whether as the primary text, a supplemental text, or a recommended resource, I ask only two small favors: (1) When you make it available to students, please always include a link back to the text's download site. While you are free to download and distribute the text under the Creative Commons 4.0 license, my preference is that you point students to this website to download it themselves. Seeing the download numbers tick up is a treat, and I plan to add additional appendices over time, so the download file will be updated occasionally. (2) Please send me a quick email at Christopher-Horne@utc.edu letting me know you're using it. I welcome your feedback as well. Thank you, and best wishes for successful research methods instruction.

This book was produced with Pressbooks (https://pressbooks.com) and rendered with Prince.

CONTENTS

Main

ABOUT THE AUTHOR

I am the Dalton Roberts Professor of Public Administration and MPA Program Coordinator at The University of Tennessee at Chattanooga, where I teach undergraduate and graduate courses in research methods, program planning, and program evaluation. I also support students learning outside the classroom through engagement in our community. I hold a Ph.D. in Public Policy from the joint program of Georgia Tech and Georgia State University. My own research has to do with interactions between public policy and the nonprofit sector and faith-work integration (or not) in public administration. In addition to being a professor, I also maintain a small program evaluation consulting practice, serve on the board of a large nonprofit organization, teach Sunday School, play a little piano, make a little art, and enjoy being a homebody with my wife and four kids.

A NOTE TO STUDENTS

A few years ago, I decided to write this summary of my research methods course on the spur of the moment, but my motives were longstanding. The prices of social science research methods textbooks are ridiculous. It's not like this is top secret knowledge mastered by only a select, highly specialized few. Really, anyone with a graduate degree in any social science discipline knows this stuff. Since writing that first version, an army of likeminded educators has assembled to develop inexpensive alternatives to traditional textbooks, and I'm happy to sign up. In the third version, I removed the word "free" from the title only because an inexpensive printed version is available for purchase on Amazon. I've observed that most students print this entire document anyway, and several students have asked about the availability of a hard copy. The free electronic version will remain available at https://scholar.utc.edu/oer/1.

Aside from indignation over textbook prices, I also want you to learn. I know that many students won't read an expensive, dry, long textbook, but I hope that many more will read a free (or cheap), brief textbook. I've made an effort to avoid being too boring, but I can't make any promises there. I'm probably not the best judge of my own boredom quotient. (But, for what it's worth, I think this is riveting stuff.) I'm convinced that different students learn different ways, and this summary provides one more way to learn. I don't think these ways-of-learning should be treated as either-or choices. I think all students will maximize their learning by reading, zealously participating in class exercises, completing course assignments, watching YouTube videos, and listening attentively to lectures.

There's a certain freedom that comes with writing something you won't charge people to read, and I have some confessions to make. I wrote this course summary somewhat quickly. This was hard for me—I'm usually a very slow, deliberate writer, editing as I go. I found I could move along pretty quickly if I wrote in a fairly breezy style, like talking to a longsuffering friend about research methods. It made writing it easier, and I hope it will make reading it easier, too. I didn't agonize too much over the structure of this summary. I find with research methods, it's hard to teach about A before B, B before C, and C before A. I did my best, but you'll see several comments like "more about that later" where I pretty much threw up my hands. Everything's related to everything else. It's one of those topics where you have to understand the whole before you understand the parts—another reason for having a brief text you can read through to get the big picture pretty quickly. And while it's written in a fairly informal, conversational style, I didn't entirely take it easy on you. There are no elaborate outlines, no "questions for review," far fewer headings and subheadings and subsubheadings than I usually prefer, a mere smattering of bullet points, and only two diagrams. Students wishing to make the most of this summary will *study it*—outlining, taking notes, writing summaries, asking questions and seeking out answers, discussing it with your classmates—all good ideas.

I worked on this revision at a time when we debate what's "fake news" and what should count as evidence when making important decisions in public affairs. Empirical research skills cannot answer all these questions, but they can help. It's my hope that many of you will go on to learn more about research methods and to conduct your own original research. Even more, I hope *all* of you will become better equipped to critically assess the information we encounter in our civic lives and to make your own well-reasoned contributions to the discourse around issues in the public sphere that are important to you.

CSH

January, 2022

A NOTE TO INSTRUCTORS

If you've made the effort to download and read this, I'm guessing I don't need to persuade you of the value in resisting unnecessarily high costs of learning materials for our students. For-profit publishers play an important role in the academic knowledge ecosystem, but pricey textbooks don't have to be the norm across all of our students' courses.

I've used this summary (and earlier versions of it) for well over a decade in several courses: undergraduate political science research methods, undergraduate and graduate program evaluation, and graduate applied research methods. In the undergraduate research methods course, this was the only textbook, which I heavily supplemented with articles, some lecture, and a lot of in-class exercises. In the other courses, this was a supplemental text or the basis of a self-guided review. When used alongside other texts, I've found it helpful to point out that methodologists don't always use the same terms in exactly the same way (*content validity* and *construct validity* are good examples). I use this as an opportunity to talk about the social nature of research—nothing we do is in a social vacuum. Research is always done in dialogue with others, and part of that is negotiating the language we use. Generally, I think this text gets the job done, and it works well for mostly or entirely "flipped" courses. My students actually read it, perhaps more often and with less coercion than the typical longer text. I usually encourage students to read the whole thing through once, and then again, more slowly, in preparation for working with the ideas in class. I've had particular delight in former students asking for a copy of this text so they could brush up on research methods for graduate school and professional assignments—that's quite a nice reward for the work represented here.

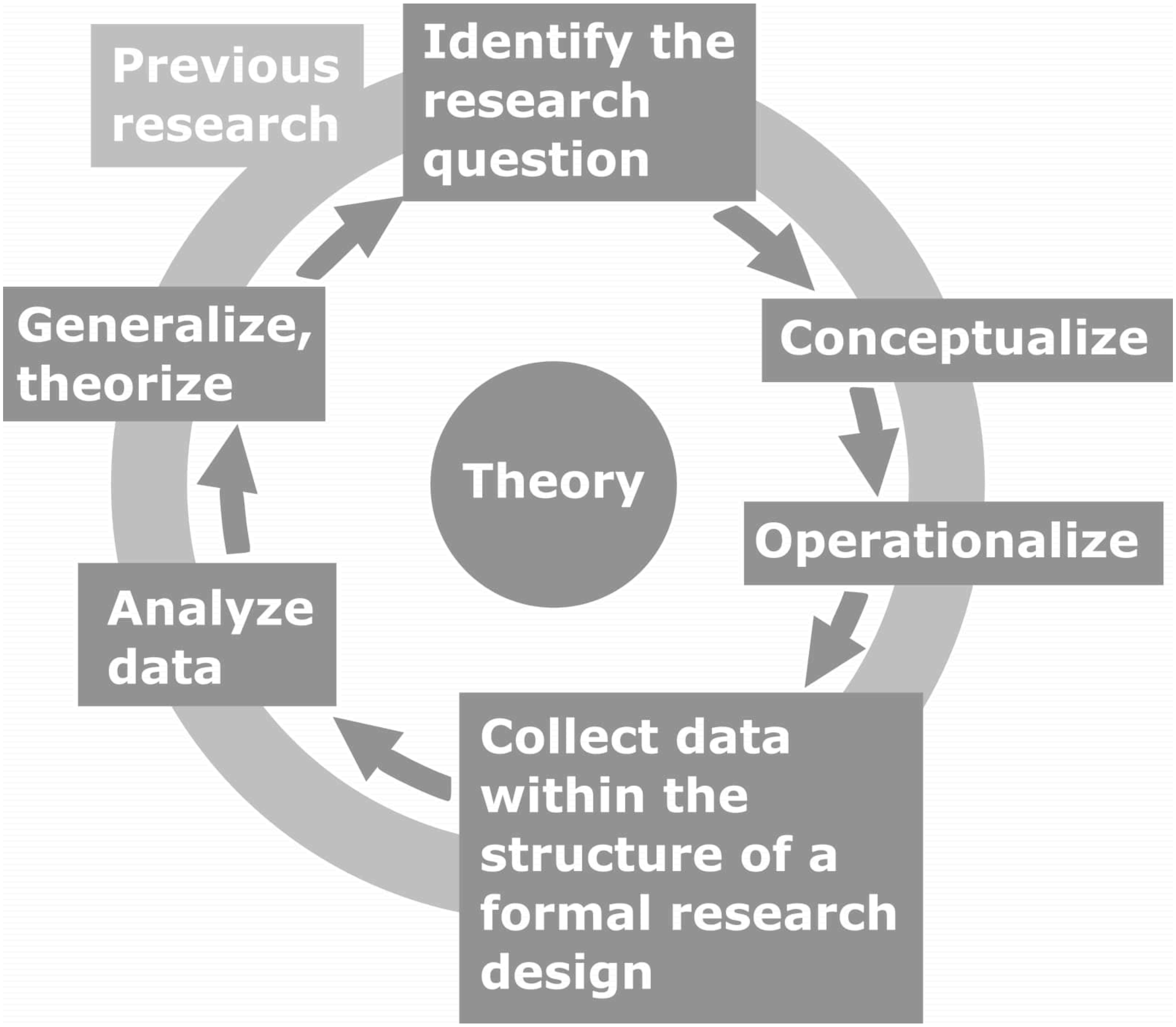
If you use this text in any way, whether as the primary text, a supplemental text, or a recommended resource, I ask only two small favors: (1) When you make it available to students, please always include a link back to the text's download site, https://scholar.utc.edu/oer/1. While you are free to download and distribute the text under the Creative Commons 4.0 license, my preference is that you point students to this website to download it themselves. Seeing the download numbers tick up is a treat, and I plan to add additional appendices over time, so the download file will be updated occasionally. (2) Please send me a quick email at Christopher-Horne@utc.edu letting me know you're using it. I certainly welcome your feedback as well. Many of the improvements to this fourth version are based on feedback I've received from instructors and students around the globe, for which I am very grateful.

Thank you, and best wishes for successful research methods instruction.
CSH

INTRODUCTION AND OUR MODEL OF THE RESEARCH PROCESS

Social science research methods are those skills and techniques we use to build knowledge about social phenomena. In this text, we are specifically interested in *empirical* social science research methods as a way of building knowledge. When using *empirical* methods, we are building knowledge based on systematic observations. Other forms of building knowledge, such as legal analysis, philosophical reasoning, and theory-building, are very important in the social sciences; they're just not the focus here.

Just like much of the social phenomena we learn about, the process of doing social research can be depicted by a model. A model, of course, is a simplification of reality and shouldn't be mistaken for the real thing (an error called *reification*). The reality is more complex and more iterative than the model suggests. It is, though, a good way to structure our thinking about the research process. Here's the model I prefer, adapted from Edward Olson and Laurence Jones's 1996 textbook, *Political Science Research*:


Previous research
Identify the research question
Conceptualize
Operationalize
Collect data within the structure of a formal research design
Analyze data
Generalize, theorize
Theory

IDENTIFYING THE RESEARCH QUESTION (AND AN ASIDE ABOUT THEORY)

The model presents the research process as circular, but *identifying the research question* is a good starting point. In this step, we specify what it is that we want to learn more about. Usually, but not always, this takes the form of a question. It could also be a statement of research purpose, though. When doing empirical research, it's important to develop a question that can be answered—or that one can attempt to answer—based on observations. A simple research question would be *How many candidates for public office use negative campaign advertisements to detract from their opponents?* We could come up with a defensible answer (we rarely come up with absolutely conclusive answers in social research) to this question based on observations.

There are other types of questions that empirical social research cannot answer. Empirical social research methods do not answer *normative* questions. Normative questions are questions that are answered based on opinions, values, and subjective preferences. Normative questions often have the word *should* in them: *Should candidates for public office use negative campaign advertisements? Should donations to churches be tax deductible? Should corporations be required to disclose lobbying expenses? Should universities consider race in making admissions decisions?* In these examples, no amount of systematic observation can provide a defensible answer to the question; ultimately, answering these questions is a matter of subjective values. However—and this is a very important however—empirical research can help us develop better informed opinions about these normative questions. To help develop a better informed opinion about whether or not candidates should use negative campaign ads, a researcher might investigate related empirical questions, such as *How do negative campaign ads affect voter behavior?* and *How do negative campaign ads affect voters' opinions about the endorsing candidate?* Social researchers, then, don't run away from normative questions—most interesting questions are normative—but, instead, look for opportunities for empirical research to shed light on normative questions.

Even this, though, is oversimplifying a bit too much. It's naïve to think that doing empirical research is value-free. Our values influence our decisions throughout the entire research process, from what we study, to how we make observations, to how we make sense of what we observe. Objectivity is a worthy goal when doing empirical social research, but it is an elusive goal, and we should always try to be aware of and transparent about how our own biases affect our research.

Still other interesting questions are the domain of legal analysis, philosophy, or history, not empirical social science research. Legal analysis is required to tackle questions like *Can state governments constitutionally cede authority to local governments to allow or ban carrying handguns in public parks?* Questions about events from the distant past (an admittedly ambiguous standard) are generally left to historians, though some questions reside in a gray area where empirical research methods could be used to learn about historical events.

The distinction between the domains of social research and history raises an important point: When conducting social research, our goal is usually to build knowledge that is *generalizable*; that is, we usually want to be able to apply what we learned from our observations to other cases, settings, or times. We may make observations of one local election, but with the goal of generating knowledge that could be applied to local elections in other jurisdictions, to future or past local elections, or to citizen participation in administrative rulemaking at the local level. While historians may be more likely to do research to build in-depth knowledge about a single case, we rarely undertake a social research project with the goal of generating knowledge that would be applied *only* to understanding what we've directly observed. (A partial exception to this would be when we conduct case studies, discussed later—but this is only a partial exception.)

Empirical research questions can have different purposes. Some empirical social science research questions seek to *describe* social phenomena. Sometimes, you'll see the phrase *mere description* used, and some research methods textbook authors will say that description doesn't even count as research. This is nonsense. Describing social phenomena based on systematic observations is certainly a legitimate purpose of social science research.

When these textbook authors diminish the importance of description, what they have in mind as more suitable research purposes are *explanation* and *prediction*. By pursuing these research purposes, we are now exploring questions of causality. If we're *explaining* something, we've observed something occur, and then we're looking back in time, in a sense, to figure out what caused it to occur: *Why were high- and middle-income independent voters less likely to vote for the Democratic candidate than low-income independent voters in the last gubernatorial election?* There, we've observed something interesting about the last gubernatorial election, and we want to figure out what happened before to *explain* it. If we're *predicting* something, we observe past trends or the state of things now and use those observations to predict what will happen in the future: *How will low-income voters vote in the upcoming state senate election?* We'll come back to the notion of causality shortly.

Research questions with the purposes of description, explanation, and prediction are all pursued using a broad range of social research methods. A fourth research purpose, *understanding*, though, is more tightly coupled with a narrower range of research methods—those methods that center around collecting and analyzing *qualitative data*. *Qualitative data* are usually words, but they can also be pictures or sounds—basically, any data that are not numeric. Transcripts of interviews with campaign managers, the text of administrative agencies' requests for proposals, the text of Supreme Court opinions, survey respondents' answers to open-ended questions, and pictures of people in a political protest are all examples of qualitative data. (Quantitative data, on the other hand, are numeric. More on different types of data later.) With the research purpose of *understanding*, we are not using the term "understanding" in its colloquial sense; instead, we mean "understanding" with the connotation of *verstehen*, a German word that doesn't translate into English very well but carries the idea of understanding someone else's subjective experiences. When conducting research with the goal of *verstehen*, we want to achieve an in-depth understanding of others' opinions, attitudes, motivations, beliefs, conceptual maps, and so on. Typically, this would involve talking with them, listening to their words, or reading what they've written—thus the association of qualitative data collection with research questions that have the goal of achieving understanding-qua-*verstehen*.

To be clear: Research projects with the purposes of description, explanation, and prediction use the full range of research methods, including the collection of both quantitative and qualitative data; research projects with the purpose of understanding generally use methods focused on collecting qualitative data.

Research questions, then, can pursue one or more of these four purposes—description, explanation, prediction, and understanding—but where do research questions come from? At some point in their studies, most students will know the fear of the blank page: Where do I start? What is my research question? Research questions might occasionally arrive in a flash of inspiration, but, usually, their origins are more mundane and require more work. I think most social researchers would agree that their research questions come from some combination of four starting points: deduction, induction, previous research, and what I'll just describe for now as one of the research profession's dirty little secrets.

The classic "correct" textbook answer to the question of where research questions come from is deduction from theory. By employing deductive thinking, we start with a theory and deduce the research questions that it suggests.

Before going any further into deducing research questions, though, we should pause for a moment on that other term, *theory*. A theory is simply a set of concepts and relationships among those concepts that helps us understand or explain some phenomenon—for us, a social phenomenon. Sometimes, theories are very formal; they're written down in a concise statement in a definitive form by a specific author or group of authors, and they include a wholly specified set of concepts; everybody knows what's in the theory and what's out. Maslow's Hierarchy of Needs—that model of human motivation that crops up in every other undergraduate course— comes to mind as an example of a formal theory. In this theory, a specific set of concepts (the need for socialization, the need for security, and so on) are related in a specific way to explain why people do what they do. Other theories, though, are relatively loose; they're evolving, they're gleaned from across a wide range of writings and assembled in different ways by different people, and there might be disagreement over precisely which concepts are included and which are not. I once used something called "crowding out theory" as it applies to charitable giving to nonprofit organizations, and I had to piece together my own version of this theory by reading what a lot of other people had written about it. My version would have looked somewhat like others', but not identical. My formulation of the theory linked concepts like charitable giving, government funding, donors' perceptions of government funding, and nonprofit managerial capacity to predict how charitable donors would react to nonprofit organizations receiving different types of government subsidy.

(A quick aside to students interested in studying public policies, programs, and organizations. *You are my people.* When we conduct research about a particular program, public policy, or organization, a model of the program, policy, or organization often plays the role of theory in the research process. A logic model, for example, depicts a program in terms of its inputs, activities, outputs, and outcomes—not unlike a set of concepts and relationships among those concepts. I've provided an example of a logic model and how it can generate a lot of applied research questions in Appendix A.)

... Everyone else—just in case you skipped that last paragraph: You should read Appendix A, too—you'll find the examples of empirical research questions helpful.

A theory (or program, policy, or organization model), needn't be such a complicated thing, but I think many students are like I was as an undergraduate student (and even into my graduate student years): intimidated by theory. I didn't totally understand what theory was, and I thought handling theory was best left to the professionals. Like most students, I thought of theory as an antique car—the kind of antique car that is kept in pristine condition, all shiny and perfect, in its climate-controlled garage, rolled out only to show off, and then rolled back in for safe keeping. It turns out, though, that most researchers don't view theories this way at all. Instead, they view their theories as beat-up pickup trucks. They're good insofar as they're useful for doing their job. It's OK if they get dinged up in the process. They're not just rolled out for showing off; they're used to help understand the world, driven as far as they'll go. (I stole this analogy from one of my professors, Gordon Kingsley, but, like a good theory, I've modified it a bit to suit our purposes here.)

As suggested by our model of the research process, theory is at the center of the entire process (not just at the beginning like in some other models). It's the touchstone for every step along the way, including the step at hand: identifying a research question. To develop a research question, we can start with a theory and all its concepts and relationships among those concepts to deduce research questions—questions that, essentially, ask whether the theory matches observations in the real world. Maslow's Hierarchy of Needs, for example, might suggest the question, *Are voters whose basic needs are not being met more likely than others to support candidates who promise to alleviate citizens' security and safety needs?* Here, we have developed a question that uses a theory as a starting point for explaining a political phenomenon. How did we deduce this research question from our theory? The theory helped us identify relevant concepts, like voters' security and safety needs and candidates' promises to alleviate them, and a potential relationship between these concepts and what we're interested in explaining, voters' choice of candidate. (And like most empirical research based on Maslow's theory, alas, we might have difficulty finding much empirical support for it.)

Research questions may also be developed *inductively* by observing social phenomena and then developing research questions based on what has been observed. Perhaps you observe more men than women in your political science courses but not in your other courses. You can make this casual observation the basis of a research question: *Are men more likely to take political science courses than women?* or *How does students' sex relate to their course selection?* or *How does gender socialization affect students' selection of majors?* Researchers with an application orientation may simply experience a problem and develop a research question to figure out how to overcome it: *Why did unemployment benefit claim processing time increase by 50% last year?* You may find that your casual observations reflect regularities confirmed through systematic observations, and, ultimately, you may even develop a theory or modify an existing theory based on what you learned. So, whereas a deductive research process begins with theory and generalizations that lead to observation, an inductive research process begins with observations that lead to generalizations and theory.

Our model of the research process points to another source of research questions: previous research. *Previous research* usually refers to all of the publications that report the results of research that has already been conducted on a given topic. We use previous research to develop research questions in a couple of ways. If there's a social phenomenon we'd like to learn more about, a good starting point is to read all of the previous research on that topic. Once we have a command of that body of knowledge, we can identify gaps, internal inconsistencies, unresolved questions, and emerging research directions in the literature. It's one small step further to develop research questions that build on the existing body of research. Sometimes, using previous research is more literal; often, an article, chapter, or book will include a section titled something like "Recommendations for future research," and, voilà, you have a research question. (As portrayed in the model, generating research questions isn't the only use of previous research; it's used throughout the entire research process, as we'll see.)

And then there's the dirty little secret of the social research professions. Sometimes we begin, not nobly with a theory, not astutely with our own observations, not studiously with previous research, but shamelessly with available data. An aspiring researcher can simply comb through data in hand in search of a research question that can be asked of it. Have access to data collected through the General Social Survey, a public opinion survey conducted every two years?

Read through the table of contents, find some questions that might go together, and try it out. Let the availability of the data—not theoretic or practical import or even your own casual observations—make you interested in a research question. This approach is roundly criticized because it smacks of data fishing; it's almost always possible to find *some* patterns in your data, even if it's just a fluke. Data fishing is exploiting these fluky patterns by making them seem important even when they're not. Baseless dataset dredging is not a good starting point for conducting research. It happens, though. Untenured assistant professors and dissertation-writing doctoral students are under tremendous pressure to publish research, and the unfortunate truth is that papers reporting "null findings" don't get published very often. Safer to start with a pattern you've stumbled upon in your data and then figure out how to make it sound important, like something you went looking for, so the thinking goes. This approach isn't entirely bad; there are legitimate ways to conduct data mining (the more acceptable term). Data are collected because *someone* thought they were important, so it's not inconceivable that you could uncover important, unanticipated patterns in your data. Thinly disguised data fishing, though, is quickly identified and disregarded by other scholars.

Before we wrap up our consideration of research questions, we should spend a moment unpacking the notion of causality. Three concepts will help us understand how social research approaches questions of cause-and-effect: *probabilistic causality*, *multiple causation*, and *underlying causal mechanisms*. When we seek causal explanations in social research, we rarely talk in absolutes. The type of causality often studied in the physical sciences is *deterministic causality*, meaning definite cause-and-effect relationships: *Flipping the switch causes the light to come on*. In the social sciences (though not exclusively in the social sciences), we are almost always studying questions of probabilistic causality, meaning cause-and-effect relationships that are more or less likely to occur: *People are less likely to vote for incumbents when the unemployment rate is high*. We are also almost always explaining and predicting phenomena that have multiple, interacting causes—multiple causation. Why do some people have higher incomes than others? This surely has many causes—education, age, ability, parents' wealth, motivation, discrimination, opportunity, job choice, attitudes toward work and money, and so on. And these causes, themselves, affect each other. Much advanced social research attempts to figure out these complex, interacting cause-and-effect relationships. When we make causal claims like *age affects income*, we are really masking a more complex web of cause-and-effect relationships. Does our age really, inherently, affect our income? Not really. Age affects income in the sense that this ostensible relationship is the manifestation of a more complex underlying causal mechanism. This underlying causal mechanism explains *why* age seems to affect income—a cause-and-effect story about biological development, the accumulation of education and experience, and the demands of different stages of life. We'll revisit underlying causal mechanisms in the next section when we learn about independent and dependent variables.

CONCEPTUALIZING AND OPERATIONALIZING (AND SOMETIMES HYPOTHESIZING)

Research questions are an essential starting point, but they tend to be too abstract. If we're ultimately about making observations, we need to know more specifically what to observe. *Conceptualization* is a step in that direction. In this stage of the research process, we specify what concepts and what relationships among those concepts we need to observe. My research question might be *How does government funding affect nonprofit organizations?* This is fine, but I need to identify what I want to observe much more specifically. Theory (like the crowding out theory I referred to before) and previous research help me identify a set of concepts that I need to consider: different types of government funding, the amount of funding, effects on fundraising, effects on operations management, managerial capacity, donor attitudes, policies of intermediary funding agencies, and so on. It's helpful at this stage to write what are called *nominal definitions* of the concepts that are central to my study. These are definitions like what you'd find in a dictionary, but tailored to your study; a nominal definition of *government subsidy* would describe what I mean in this study when I use the term.

After identifying and defining concepts, we're ready to *operationalize* them. To operationalize a concept is to describe how to measure it. (Some authors refer to this as the *operational definition*, which I find confuses students since it doesn't necessarily look like a definition.) Operationalization is where we get quite concrete: To operationalize the concept *revenue of a nonprofit organization*, we might record the dollar amount entered in line 12 of their most recent Form 990 (a financial statement nonprofit organizations must file with the IRS annually). This dollar amount will be my *measure* of nonprofit revenue.

Sometimes, the way we operationalize a concept is more indirect. *Public support* for nonprofit organizations, for example, is more of a challenge to operationalize. We might write a nominal definition for *public support* that describes it as having something to do with the sum of individuals' active, tangible support of a nonprofit organization's mission. We might operationalize this concept by recording the amount of direct charitable contributions, indirect charitable contributions, revenue from fundraising events, and the number of volunteer hours entered in the respective Form 990 lines.

Note that when we operationalized nonprofit revenue, the operationalization yielded a single measure. When we operationalized public support, however, the operationalization yielded multiple measures. Public support is a broader, more complex concept, and it's hard to think of just one measure that would convincingly represent it. Also, when we're using measures that measure the concept more indirectly, like our measures for public support, we'll sometimes use the word *indicator* instead of *measure*. The term *indicator* can be more accurate. We know that measuring something as abstract as public support would be impossible; it is, after all, a social construct, not something concrete. Our measures, then, *indicate* the level of public support more than actually *measure* it.

I just slipped in that term, *social construct*, so we should go ahead and face an issue we've been sidestepping so far: Many concepts we're interested in aren't observable in the sense that they can't be seen, felt, heard, tasted, or smelled. But aren't we supposed to be building knowledge based on observations? Are unobservable concepts off limits for empirical social researchers? Let's hope not! Lots of important concepts (maybe all the most important concepts) are social constructs, terms that don't have meaning apart from the meaning we, collectively, assign to them. Consider political literacy, racial prejudice, voter intent, employee motivation, issue saliency, self-esteem, managerial capacity, fundraising effectiveness, introversion, and Constitutional ideology. These terms are a shorthand for sets of characteristics that we all more or less agree "belong" to the concepts they name. Can we observe political ideology? Not directly, but we can pretty much agree on what observations serve as indicators for political ideology. We can observe behaviors, like putting bumper stickers on cars, we can see how people respond to survey items, and we can hear how people respond to interview questions. We know we're not directly measuring political ideology (which is impossible, after all, since it's a social construct), but we can persuade each other that our measures of political ideology make sense (which seems fitting, since, again, it's a social construct).

Each indicator or measure—each observation we repeat over and over again—yields a *variable*. The term *variable* is one of these terms that's easier to learn by example than by definition. The definition, though, is something like "a logical grouping of attributes." (Not very helpful!) Think of the various attributes that could be used to describe you and your friends: brown hair, green eyes, 6'2" tall, brown eyes, black hair, 19 years old, 5'8" tall, blue eyes, and so on. Obviously, some of these attributes go together, like green eyes, brown eyes, and blue eyes. We can group these attributes together and give them a label: eye color. Eye color, then, is a variable. In this example, the variable eye color takes on the *values* green, brown, and blue. In many research designs, our goal in making observations is to assign *values* to *variables* for *cases*. *Cases* are the things—here, you and your friends—that we're observing and to which we're assigning values. In social science research, cases are often individuals (like individual voters or individual respondents to a survey) or groups of people (like families or organizations), but cases can also be court rulings, elections, states, committee meetings, and an infinite number of other things that can be observed. The term *unit of analysis* is used to describe cases, too, but it's a more general term; if your cases are firefighters, then your unit of analysis is the individual.

Getting this terminology—cases, variables, values—is essential. Here are some examples of cases, variables, and values . . .

- *Cases*: undergraduate college students; *variable*: classification; *values*: Freshmen, Sophomore, Junior, Senior;
- *Cases*: states; *variable*: whether or not citizen referenda are permitted; *values*: yes, no;
- *Cases*: counties; *variable*: type of voting equipment; *values*: manual mark, punch card, optical scan, electronic;

- *Cases*: clients; *variable*: length of time it took them to see a counselor; *values*: any number of minutes;
- *Cases*: Supreme Court dissenting opinions; *variable*: number of signatories; *values*: a number from 0 to 4;
- *Cases*: criminology majors; *variable*: GPA; *values*: any number from 0 to 4.0.

Researchers have a language for describing variables. A variable's *level of measurement* describes the structure of the values it can take on, whether nominal, ordinal, interval, or ratio. Nominal and ordinal variables are the *categorical variables*; their values divide up cases into distinct categories. The values of *nominal-level variables* have no inherent order. The variable sex can take on the values male and female; eye color—brown, blue, and green eyes; major— political science, sociology, biology, etc. Placing these values in one order—brown, blue, green— makes just as much sense as any other—blue, green, brown. The values of *ordinal-level variables*, though, have an inherent order. Classification—freshmen, sophomore, junior, senior; love of research methods—low, medium, high; class rank—first, second, . . . , 998th. These values can be placed in an order that makes sense—first to last (or last to first), least to most, best to worst, and so on. A point of confusion to be avoided: When we collect and record data, sometimes we assign numbers to values of categorical variables (like brown hair equals 1), but that's just for the sake of convenience. Those numbers are just placeholders for the actual values, which remain categorical.

When values take on actual numeric values, the variables they belong to are *numeric variables*. If a numeric variable takes on the value *28*, it means there are actually 28 of something—28 degrees, 28 votes, 28 pounds, 28 percentage points. It makes sense to add and subtract these values. If one state has a 12% unemployment rate, that's 3 more points than a state with a 9% unemployment rate. Numeric variables can be either interval-level variables or ratio-level variables. When *ratio-level variables* take on the value zero, zero means zero—it means *nothing* of whatever we're measuring. Zero votes means no votes; zero senators means no senators. Most numeric variables we use in social research are ratio-level. (Note that many ratio-level variables, like height, age, states' number of senators, would never actually take on the value zero, but if they did, zero would mean zero.) Occasionally, zero means something else besides nothing of something, and variables that take on these odd zeroes are interval-level variables. Zero degrees means—well, not "no degrees," which doesn't make sense. Year zero doesn't mean the year that wasn't. We can add and subtract the values of interval-level variables, but we cannot multiply and divide them. Someone born in 996 is not half the age of someone born in 1992, and 90 degrees is not twice as hot as 45.

We can sometimes choose the level of measurement when constructing a variable. We could measure age with a ratio-level variable (the number of times you've gone around the sun) or with an ordinal-level variable (check whether you're 0-10, 11-20, 21-30, or over 30). We should make this choice intentionally because it will determine what kinds of statistical analysis we can do with our data later. If our data are ratio-level, we can do any statistical analysis we want, but our choices are more limited with interval-level data, still more limited with ordinal-level data, and most limited with nominal-level data. (See Appendix E on equity in research for an explanation of how *dummy coding* can be used to helpfully transform categorical variables to ratio-level variables.)

Variables can also be described as being either *continuous* or *discrete*. Just like with the level of measurement, we look at the variable's values to determine whether it's a continuous or discrete variable. All categorical variables are discrete, meaning their variables can only take on specific, discrete values. This is in contrast to some (but not all!) numeric variables. Take temperature, for example. For any two values of the variable *temperature*, we can always imagine a case with a value in between them. If Monday's high is 62.5 degrees and Tuesday's high is 63.0 degrees, Wednesday's high could be 62.75 degrees. Temperature, then, measured in degrees, is a continuous variable. Other numeric variables are discrete variables, though. Any variable that is a count of things is discrete. For the variable *number of siblings*, Anna has two siblings and Henry has three siblings. We cannot imagine a person with any number of siblings between two and three—nobody could have 2.5 siblings. *Number of siblings*, then, is a discrete variable. (Note: Some textbooks and websites incorrectly state that all numeric variables are continuous. Do not be misled.)

If we're engaging in causal research, we can also describe our variables in terms of their role in causal explanation. The "cause" variable is the *independent variable*. The "effect" variable is the *dependent variable.* If you're interested in determining the effect of level of education on political party identification, level of education is the independent variable, and political party identification is the dependent variable.

I'm being a bit loose in using "cause" and "effect" here. Recall the concept of underlying causal mechanism. We may identify independent and dependent variables that really represent a much more complex underlying causal mechanism. Why, for example, do people make charitable contributions? At least four studies have asked whether people are more likely to make a contribution when the person asking for it is dressed nicely. (See the examples cited in Bekkers and Wiepking's 2010 "A Literature Review of Empirical Studies of Philanthropy," *Nonprofit and Voluntary Sector Quarterly*, volume 40, p. 924, which I also recommend for its many examples of how social research explores questions of causality.) Do these researchers believe the quality of stitching might affect altruism? Sort of, but not exactly. More likely, they believe potential donors' perceptions of charitable solicitors may shape their attitudes toward the requests, which will make them more or less likely to respond positively. It's a bit reductionist to say charitable solicitors' clothing "causes" people to make charitable donations, but we still use the language of independent variables and dependent variables as labels for the quality of the solicitors' clothing and the solicitees' likelihood of making charitable donations, respectively. Think carefully about how this might apply anytime an independent variable—sometimes more helpfully called an explanatory variable—is a demographic characteristic. Women, on average, make lower salaries than men. Does sex "cause" salary? Not exactly, though we would rightly label sex as an independent variable and salary as a dependent variable. Underlying this simple dyad of variables is a set of complex, interacting, causal factors—gender socialization, discrimination, occupational preferences, economic systems' valuing of different jobs, family leave policies, time in labor market—that more fully explain this causal relationship.

Identifying independent variables (IVs) and dependent variables (DVs) is often challenging for students at first. If you're unsure which is which, try plugging your variables into the following phrases to see what makes sense:

- IV causes DV
- Change in IV causes change in DV
- IV affects DV
- DV is partially determined by IV
- A change in IV predicts a change in DV
- DV can be partially explained by IV
- DV depends on IV

In the later section on formal research designs, we'll learn about control variables, another type of variable in causal studies often used in conjunction with independent and dependent variables.

Sometimes, especially if we're collecting quantitative data and planning to conduct inferential statistical analysis, we'll specify hypotheses at this point in the research process as well. A *hypothesis* is a statement of the expected relationship between two or more variables. Like operationalizing a concept, constructing a hypothesis requires getting specific. A good hypothesis will not just predict that two (or more) variables are related, but how. So, not *Political science majors' amount of volunteer experience will be related to their choice of courses,* but *Political science majors with more volunteer experience will be more likely to enroll in the public policy, public administration, and nonprofit management courses*. Note that you may have to infer the actual variables; hypotheses often refer only to specific values of the variables. Here, *public policy, public administration,* and *nonprofit management courses* are values of the implied variable, *types of courses*.

DATA COLLECTION STRUCTURED BY FORMAL RESEARCH DESIGNS

Data collection is the act of making and recording systematic observations. Those records of our observations become our *data*. The decisions facing the researcher embarking on data collection are myriad: What or who will your cases be? What kind of data will you collect? How will you structure your data collection so that you can convincingly draw conclusions from it later?

Sampling

The selection of cases to observe is the task of *sampling*. If you're going to be collecting data from people, you might be able to talk to every person that you want your research to apply to, that is, your *population*. If you're doing a study of state election commissioners, you might be able to talk to all 50 of them. In that case, you'd be conducting a *census* study. Often, though, we're only able to collect data from a portion of the population, or a *sample*. We devise a *sampling frame*, a list of cases we select our sample from—ideally, a list of all cases in the population—but then which cases do we select for the sample? We select cases for our sample by following a *sampling design*, which comes in two basic varieties: probability sampling designs and nonprobability sampling designs.

In *probability sampling designs*, every case in the population has a known, greater-than-zero probability of being selected for the sample. This feature of probability sampling designs, along with the wonder of the central limit theorem and law of large numbers, allows us to do something incredibly powerful. If we're collecting quantitative data from our sample, we can use these data to calculate *statistics*—quantified summaries of characteristics of the sample, like the median of a variable or the correlation between two variables. If we've followed a probability sampling design, we can then use statistics to estimate the *parameters*—the corresponding quantified characteristics of the population—with known levels of confidence and accuracy. This is what's going on when you read survey results in the newspaper: "± 3 points at 95% confidence." For example, if 30% of people in our sample say they'd like to work for government, then we'd be confident that if we were to repeat this survey a thousand times, 95% of the time (our level of confidence), we'd find that between 27 and 33% (because ± 3 points is our degree of accuracy) of the respondents would answer the same way. Put another way, we'd be 95% certain that 27 to 33% of the population would like to work for government.

Again, this trick of using sample statistics to estimate population parameters with known levels of confidence and accuracy only works when we've followed a probability sampling design. The most basic kind of probability sampling design is a *simple random sample*. In this design, each case in the population has a known and *equal* probability of being selected for the sample. When social researchers use the term *random*, we don't mean haphazard. (This word has become corrupted since I was in college, when my future sister-in-law started saying stuff like "A boy I knew in kindergarten just called—that was so random!" and "I just saw that guy from 'Saved by the Bell' at the mall—pretty random!") It takes a plan to be random, to give every case in the population an equal chance of being selected for a sample. If we were going to randomly select 20 state capitals, we wouldn't just select the first 20 working from west to east or the first 20 we could think of—that would introduce *sampling bias*. (We'll have more to say about *bias* later, but you get the gist of it for now.) To ensure all 50 capitals had an equal probability of being selected (a probability of 0.4, in fact), we could list them all out on a spreadsheet, use a random number generator to assign them all random numbers, sort them by those numbers, and select the first 20; or we could write each capital's name on same-sized pieces of paper, put them in a bag, shake them up, and pull out 20 names. (Some textbooks still have random number tables in the back, which you're welcome to learn how to use on your own, but they've become pretty obsolete.)

Selecting a simple random sample may be too much of a hassle because you just have a long, written list in front of you as your sampling frame, like a printed phonebook. Or, selecting a simple random sample may be impossible because you're selecting from a hypothetically infinite number of cases, like the vehicles going through an intersection. In such scenarios, you can approximate a random sample by selecting every 10th or 20th or 200th or whateverth case to reach your desired sample size, which is called *systematic sampling*. This works fine as long as *periodicity* isn't present in your population, meaning that there's nothing odd about every 10th (or whateverth) case. If you were sampling evenings to observe college life, you wouldn't want to select every 7th case, or you'd introduce severe sampling bias. Just imagine trying to describe campus nightlife by observing only Sunday evenings or only Thursday evenings. As long as periodicity isn't a problem, though, systematic sampling approximates simple random sampling.

Our goal in selecting a random (or systematic) sample is to construct a sample that is like the population so that we can use what we learn about the sample to generalize to the population. What if we already know something about our population, though? How can we make use of that knowledge when constructing our sample? We can replicate known characteristics of a sample by following another probability sampling design, a *proportionate stratified sampling design*. Perhaps we'd like to sample students at a particular college, and we already know students' sex, in-state versus out-of-state residency, and undergraduate versus graduate classification. We can use sex, residency, and classification as our *strata* and select a sample with the same proportions of male versus female, in-state versus out-of-state, and undergraduate versus graduate students as the population. If we determine that 4% of our population are male graduate students from out-of-state and we wanted a sample of 300 students, we'd select (using random sampling or systematic sampling) 12 (300*4%) male graduate students from out-of-state to be in our sample. We'd carry on similarly sampling students with other combinations of these characteristics until we had a sample proportionally representative of the population in terms of sex, residency, and classification. We probably would have gotten similar results if we had used a simple random sampling strategy, but now we've ensured proportionality with regard to these characteristics.

Sometimes, though, proportionality is exactly what we don't want. What if we were interested in comparing the experiences of students who had been homeschooled to students who were not homeschooled? If we followed a simple random sampling design or a proportionate stratified sampling design, we would probably end up with very few former homeschoolers—not enough to provide a basis of comparison to the never homeschooled. We may even want half of our sample to be former homeschoolers, which would require *oversampling* from this group to have their representation in the sample disproportionately high compared to the population, achieved by following a *disproportionate stratified sampling design*. Importantly, this is still a probability sampling design. With some careful math, we can still calculate the probability of any one case in the population being selected for the sample; it's just that for former homeschoolers, that probability would be higher than for the never homeschooled. Knowing these probabilities still permits us to use statistics to estimate parameters for the entire population of students, we just have to remember to make the responses of former homeschoolers count less and the responses of the never homeschooled count more when calculating our parameter estimates. This is done using *weights*, which are based on those probabilities, in our statistical calculations.

One final probability sampling design, *cluster sampling design*, is commonly used to sample cases that are dispersed throughout a broad geographic region. Imagine the daunting task of needing to sample 2,000 parents of kindergarteners from across the United States. There is no master list of kindergarten students or their parents to serve as a sampling frame. Constructing a sampling frame by going school to school across the country would likely consume more resources than the rest of the study itself—the thought of constructing such a sampling frame is ridiculous, really. We could, though, first randomly select, say, 20 states, and then 10 counties within each of those 20 states, and then 1 school from each of those counties, and then 10 kindergartners from each of those schools. At each step, we know the probability of each state, county, school, and kid being selected for the sample, and we can use those probabilities to calculate weights, which means we can still use statistics to estimate parameters. We'll have to modify our definition for probability sampling designs just a bit, though. We *could* calculate the probability of any one case in the population being included in the study, but we don't. Being able to calculate the probabilities of selection for each *sampling unit* (states, counties, schools, kids), though, does the same job, so we still count cluster sampling designs as one of the probability sampling designs. To modify our definition of probability sampling designs, we might say that every case in the population has a known *or knowable*, greater-than-zero probability of being selected for the sample.

Using a probability sampling design is necessary, but not sufficient, if we want to use statistics to estimate parameters. We still need an adequate sample size. How do we calculate an adequate sample size? Do we, say, select 10% of the population? It would be handy to have such an easy rule of thumb, but as it turns out, the size of the population is only one factor we have to consider when determining the required sample size. (By the way, this is probably the most amazing thing you'll learn in this text.) In addition to population size, we also have to consider required level of confidence (something you decide yourself), required level of accuracy (something else you decide), and the amount of variance in the parameter (something you don't get to decide; it is what it is).

As you'd probably guess, the larger the population size, the larger the required sample size. However, the relationship between population size and required sample size is not linear (thus no rule of thumb about selecting 10% or any other percent of the population for your sample). If we have a somewhat small population, we'll need a large proportion of it in our sample. If we have a very large population, we'll need a relatively small proportion of it in our sample. In fact, once the population size goes above around 20,000, the sample size requirement hardly increases at all (thanks again to the central limit theorem and the law of large numbers).

We also have to consider how much the parameter varies. Imagine that I'm teaching a class of 40 students, and I know that everyone in the class is the same age, I just don't know what that age is. How big would my sample size need to be for me to get a very good (even perfect) statistic, the mean age of my students? Think. One! That's right, just one. My parameter, the mean age of the class, has zero variation (my students are all the same age), so I need a very small sample to calculate a very good statistic. What if, though, my students' ages were all over the place—from one of those 14-year-old child geniuses to a 90-year-old great grandmother who decided to finish her degree? I'd be very reluctant to use the mean age of a sample of 3, 4, or even 10 students to estimate the whole class's mean age. Because the population parameter varies a lot, I'd need a large sample. The rule, then: The more the population parameter varies, the more cases I need in my sample.

The astute reader should, at this point, be thinking "Wait a sec. I'm selecting a sample so I can calculate a statistic so I can estimate a parameter. How am I supposed to know how much something I don't know varies?" Good question. Usually, we don't, so we just assume the worst, that is, we assume maximum variation, which places the highest demand on sample size. When we specify the amount of variation (like when using the sample size calculators I'll say more about below), we use the percentage of one value for a parameter that takes on only two values, like responses to yes/no questions. If we wanted to play it safe and assume maximum variation in a parameter, then, we'd specify 50%; if 50% of people in a population would answer "yes" to a yes/no question, the parameter would exhibit maximum variation—it can't vary any more than a 50/50 split. Specifying 0% or 100% would be specifying no variation, and, as it may have occurred to you already, specifying 25% would be the same as specifying 75%.

Very astute readers might have another question: "You've been referring to a required sample size, but required for what? What does it mean to have a required sample size? Isn't that what we're trying to figure out?" Another good question. Given the size of the population (something you don't control) and the amount of variance in the parameter (something else you don't control), a sample size is required to be at least a certain size if we want to achieve a desired level of confidence and a desired level of accuracy, the factors you *do* control. We saw examples of accuracy and confidence previously. We might say "I am 95% percent certain [so I have a 95% confidence level] that the average age of my class is in the 19 to 21 range [so I have a ± 1 year level of accuracy]." A clumsier way to say the same thing would be "If I were to repeat this study over and over again, selecting my sample anew each time, 95% of my samples would have average ages in the range of 19 to 21."

Confidence and accuracy go together; it doesn't make sense to specify one without specifying the other. As I've emphasized, you get to decide on your levels of confidence and accuracy, but there are some conventions in social research. The confidence level is most often set at 95%, though sometimes you'll see 90% or 99%. The level of accuracy, which is usually indicated as the range of percentage point estimates, is often set at ±1%, 3%, or 5%. If you're doing applied research, you might want to relax these standards a bit. You might decide that a survey giving you ±6% at an 85% confidence level is all you can afford, but it will help you make decisions better than no survey at all.

So far, I've just said we need to "consider" these four factors—population size, parameter variation, degree of accuracy, and degree of confidence, but, really, we have to do more than just consider them, we have to plug them into a formula to calculate the required sample size. The formula isn't all that complicated, but most people take the easy route and use a sample size calculator instead, and so will we. Several good sample size calculators will pop up with a quick internet search. You enter the information and get your required sample size in moments. Playing around with these calculators is a bit mind boggling. Try it out. What would be a reasonable sample size for surveying all United States citizens? What about for all citizens of Rhode Island? What's surprising about these sample sizes? Play around with different levels of confidence, accuracy, and parameter variation. How much do small changes affect your required sample sizes?

And note the interplay of confidence and accuracy. For any given sample size, you can have different combinations of confidence and accuracy, which will have an inverse relationship—as one goes up, the other goes down. With the same sample, I could choose either to be very confident about an imprecise estimate or to be not-so-confident about a precise estimate. I can look over a class of undergraduates and predict with near certainty that their average age is between 17 and 23, or I can predict with 75% confidence that their average age is between 19 and 20.

It's important to realize what we're getting from the sample size calculator. This is the minimum sample size if we're intending to use statistics to estimate single parameters, one by one—that is, we're calculating univariate statistics. If, however, we're planning to compare any groups within our sample or conduct any bivariate or multivariate statistical analysis with your data, our sample size requirements will increase accordingly (and necessitate consulting statistics manuals).

Calculating a minimum sample size based on the desired accuracy and confidence only makes sense if we're following a probability sampling design. Sometimes, though, our goal isn't to generalize what we learn from a sample to a population; sometimes, we have other purposes for our samples and use *nonprobability sampling designs.* Maybe we're doing a trial run of our study. We just want to try out our questionnaire and get a feel for how people will respond to it, so we use a *convenience sampling design*, which is what it sounds like—sampling whatever cases are convenient. You give your questionnaire to your roommate, your mom, and whoever's waiting in line with you at the coffee shop. Usually, convenience sampling is used for *field testing* data collection instruments, but it can also be used for *exploratory research*—research intended to help orient us to a research problem, to help us figure out what concepts are important to measure, or to help us figure out where to start when we don't have a lot of previous research to build on. We know that we have to be very cautious in drawing conclusions from exploratory research based on convenience samples, but it can provide a very good starting point for more generalizable research in the future.

In other cases, it would be silly to use a probability sampling design to select your case. What if you wanted to observe people's behavior at Green Party rallies? Would you construct a sampling frame listing all the upcoming political rallies and randomly select a few, hoping to get a Green Party rally in your sample? Of course not. Sometimes we choose our sample because we want to study particular cases. We may not even describe our case selection as sampling, but when we do, this is *purposive sampling*. We can also use purposive sampling if we wish to describe typical cases, atypical cases, or cases that provide insightful contrasts. If I were studying factors associated with nonprofit organizational effectiveness, I might select organizations that seem similar but demonstrate a wide range of effectiveness to look for previously unidentified differences that might explain the variation. Purposive sampling is prominent in studies built around in-depth qualitative data, including case studies, which we'll look at in a bit.

When purposively selecting cases of interest, we should take care not to draw unwarranted conclusions from cases *selected on the dependent variable*, the taboo sampling strategy. Imagine we want to know whether local governments' spending on social media advertising encourages local tourism. Our independent variable is social media advertisement spending, and our dependent variable is the amount of tourism. If we were to adopt this taboo sampling strategy, we would identify localities that have experienced large increases in tourism. We may then, upon further investigation, learn they had all previously increased spending on social media advertising and conclude that more advertising spending leads to more tourism. Can we legitimately draw that conclusion, though? It may be that many other localities had also increased their social media advertising spending but did not see an increase in tourism; the level of spending may not affect tourism at all. It's even possible that other localities spent more on social media advertising—we do not know because we fell into the trap of selecting cases on the dependent variable.

We may wish to do probability sampling but lack the resources, potentially making a *quota sampling design* a good option. This is somewhat of a cross between convenience sampling design and the stratified sampling designs. Before, when we wanted to include 12 male out-of- state graduate students in our sample, we constructed a sampling frame and randomly selected them. We could, however, select the first 12 male out-of-state graduate students we stumble upon, survey them to meet our quota for that category of student, and then seek out students in our remaining categories. (This is what those iPad-carrying marketing researchers at the mall and in theme parks are doing—and why they'll ignore you one day and chase you down the next.) We'd still be very tentative about generalizing from this sample to the population, but we'd feel more confident than if our sample had been selected completely as a matter of convenience.

One final nonprobability sampling design is useful when cases are difficult to identify beforehand, like meth users, sex workers, or the behind-the-scenes movers-and-shakers in a city's independent music scene. What's a researcher wanting to interview such folks to do? Post signs and ask for volunteers? Probably not. She may be able to get that first interview, though, and, once that respondent trusts her, likes her, and becomes invested in her research, she might get referred to a couple more people in this population, which could lead to a few more, and so on. This is called (regrettably, I think, because I'd hate to have the term *snowball* in my serious research report) a *snowball sampling design* or (more acceptably but less popularly) a *network sampling design*, and it has been employed in a lot of fascinating research about populations we'd otherwise never know much about.

Data Collection Methods

The decision of how to select cases to observe may present a long list of options, but deciding what specific types of data to collect presents us with infinite options. It seems to me, though, that the kinds of data collection we do in empirical social research all fall in one of three broad categories: asking questions, making direct observations, and collecting secondary data.

Collecting data by asking questions can be somewhat like our everyday experience of carrying on conversations. If you have taken an introductory communications course, you have learned how interpersonal communication involves encoding our intended meaning in words, transmitting those words to our conversation partner, who then receives those words, decodes them to derive meaning, and then repeats the process in response. All of this can be derailed due to distractions, assumptions, moods, attitudes, social pressures, and motives. In normal conversation, both parties can try to keep communication on track by reading body language, asking clarifying questions, and correcting misunderstandings. When asking questions for research, though, you—the researcher—are solely responsible for crafting a question-and- answer exchange that yields valid data. The researcher must ensure the meaning she intends to encode in her questions are accurately decoded by the respondent; she must ensure the respondent is enabled to accurately encode his intended meaning in his available response options; she must anticipate and mitigate threats to the accurate encoding and decoding of meaning posed by those distractions, assumptions, moods, attitudes, social pressures, and motives. Before thinking about the nuts and bolts of asking questions for research, understand that it is, essentially, two-way communication with all responsibility for ensuring its accuracy on the head of the researcher.

Volumes have been written about the craft of asking people questions for research purposes, but we can sum up the main points briefly. Researchers ask people questions face-to-face (whether in person or via web-based video conferencing), by telephone, using self-administered written questionnaires, and in web-based surveys. Each of these *modes of administration* has its advantages and disadvantages. It's tempting to think that face-to-face interviewing is always the best option, and often, it *is* a good option. Talking to respondents face-to-face makes it hard for them to stop midway through the interview, gives them the chance to ask questions if something needs clarifying, and lets you read their body language and facial expressions so you can help if they look confused. A face-to-face interview gives you a chance to build rapport with respondents, so they're more likely to give good, thorough answers because they want to help you out. That's a double-edged sword, though: Having you staring a respondent in the face might tempt him to give answers that he thinks you want to hear or that make him seem like a nice, smart, witty guy—the problem of *social desirability bias*.

Combating bias is one of the most important tasks when designing a research project. *Bias* is any systematic distortion of findings due to the way that the research is conducted, and it takes many forms. Imagine interviewing strangers about their opinions of a particular political candidate. How might their answers be different if the candidate is African-American and the interviewer is white? What if the respondent is interviewed at her huge fancy house and the interviewer is wearing tattered shoes? The human tendencies to want to be liked, to just get along, and to avoid embarrassment are very strong, and they can strongly affect how people answer questions asked by strangers. To the extent that respondents are affected similarly from interview to interview, the way the research is being conducted has introduced bias.

So, then, asking questions face-to-face may be a good option sometimes, but it may be the inferior option if social desirability bias is a potential problem. In those situations, maybe having respondents answer questions using a self-administered written questionnaire would be better. Completing a questionnaire in private goes a long way in avoiding social desirability bias, but it introduces other problems. Mail is easier to ignore than someone knocking at your door or making an appointment to meet with you in your office. You have to count more on the respondent's own motivation to complete the questionnaire, and if motivated respondents' answers are systematically different than unmotivated nonrespondents, your research plan has introduced *self-selection bias*. You're not there to answer questions the respondent may have, which pretty much rules out complicated questionnaire design (such as questionnaires with a lot of skip patterns—"If 'Yes,' go to Question 38; if 'No,' go to Question 40" kind of stuff). On the plus side, it's much easier and cheaper to mail questionnaires to every state's director of human services than to visit them all in person.

You can think through how these various pluses and minuses would play out with surveys administered by telephone. If you're trying to talk to a representative sample of the population, though, telephone surveys have another problem. Think about everyone you know under the age of 30. How many of them have telephones—actual land lines? How many of their parents have land lines? Most telephone polling is limited to calling land lines, so you can imagine how that could introduce *sampling bias*—bias introduced when some members of the population are more likely to be included in a study than others. When cell phones are included, you can imagine that there are systematic differences between people who are likely to answer the call and those who are likely to ignore the unfamiliar Caller ID—another source of sampling bias. If you are a counseling center administrator calling all of your clients, this may not be a problem; if you are calling a randomly selected sample of the general population, the bias could be severe.

Web-based surveys have become a very appealing option for researchers. They are incredibly cheap, allow complex skip patterns to be carried out unbeknownst to respondents, face no geographic boundaries, and automate many otherwise tedious and error-prone data entry tasks. For some populations, this is a great option. I once conducted a survey of other professors, a population with nearly universal internet access. For other populations, though—low-income persons, homeless persons, disabled persons, the elderly, and young children—web-based surveys are often unrealistic.

Deciding what medium to use when asking questions is probably easier than deciding what wording to use. Crafting useful questions and combining them into a useful data collection instrument take time and attention to details easily overlooked by novice researchers. Sadly, plentiful examples of truly horribly designed surveys are easy to come by. Well-crafted questions elicit unbiased responses that are useful for answering research questions; poorly crafted questions do not.

So, what can we do to make sure we're asking useful questions? There are many good textbooks and manuals devoted to just this topic, and you should definitely consult one if you're going to tackle this kind of research project yourself. Tips for designing good data collection instruments for asking questions, whether questionnaires, web-based surveys, interview schedules, or focus group protocols, boil down to a few basics.

Perhaps most important is paying careful attention to the wording of the questions themselves. Let's assume that respondents want to give us accurate, honest answers. For them to do this, we need to word questions so that respondents will interpret them in the way we want them to, so we have to avoid ambiguous language. (What does *often* mean? What is *sometimes*?) If we're providing the answer choices for them, we also have to provide a way for respondents to answer accurately and honestly. I bet you've taken a survey and gotten frustrated that you couldn't answer the way you wanted to.

I was once asked to take a survey about teaching online. One of the questions went something like this:

Do you think teaching online is as good as teaching face-to-face?

❑ *Yes*
❑ *No*
❑ *I think they're about the same*

I've taught online lot, I've read a lot about online pedagogy, I've participated in training about teaching online, and this was a frustrating question for me. Why? Well, if I answer *no*, my guess is that the researchers would infer that I think online teaching is inferior to face-to-face teaching. What if I am an online teaching zealot? By *no*, I may mean that I think online teaching is superior to face-to-face! There's a huge potential for disconnect between the meaning the respondent attaches to this answer and the meaning the researcher attaches to it. That's my main problem with this question, but it's not the only one. What is meant, exactly, by *as good as*? *As good as* in terms of what? In terms of student learning? For transmitting knowledge? My own convenience? My students' convenience? A respondent could attach any of these meanings to that phrase, regardless of what the researcher has in mind. Even if I ignore this, I don't have the option of giving the answer I want to—the answer that most accurately represents my opinion—*it depends*. What conclusions could the researcher draw from responses to this question? Not much, but uncritical researchers would probably report the results as filtered through their own preconceptions about the meanings of the question and answer wording, introducing a pernicious sort of bias—difficult to detect, particularly if you're just casually reading a report based on this study, and distorting the findings so much as to actually convey the opposite of what respondents intended. (I was so frustrated by this question and fearful of the misguided decisions that could be based on it that I contacted the researcher, who agreed and graciously issued a revised survey—research methods saves the day!)

Question wording must facilitate unambiguous, fully accurate communication between the researcher and respondent.

Just as with mode of administration, question wording can also introduce social desirability bias. Leading questions are the most obvious culprit. A question like *Don't you think public school teachers are underpaid?* makes you almost fall over yourself to say "Yes!" A less leading question would be *Do you think public school teachers are paid too much, paid too little, or paid about the right amount?* To the ear of someone who doesn't want to give a bad impression by saying the "wrong" answer, all of the answers sound acceptable. If we're particularly worried about potential social desirability bias, we can use *normalizing statements*: *Some people like to follow politics closely and others aren't as interested in politics. How closely do you like to follow politics?* would probably get fewer trying-to-sound-like-a-good-citizen responses than *Do you stay well informed about politics?*

Closed-ended questions—questions that give answers for respondents to select from—are susceptible to another form of bias, *response set bias*. When respondents look at a range of choices, there's subconscious pressure to select the "normal" response. Imagine if I were to survey my students, asking them:

How many hours per week do you study?

- ❑ *Less than 10*
- ❑ *10 – 20*
- ❑ *More than 20*

That middle category just looks like it's the "normal" answer, doesn't it? The respondent's subconscious whispers "Lazy students must study less than 10 hours per week; more than 20 must be excessive." This pressure is hard to avoid completely, but we can minimize the bias by anticipating this problem and constructing response sets that represent a reasonable distribution.

Response sets must be exhaustive—be sure you offer the full range of possible answers—and the responses must be mutually exclusive. How *not* to write a response set:

How often do you use public transportation?

- ❑ *Never*
- ❑ *Every day*
- ❑ *Several times per week*
- ❑ *5 – 6 times per week*
- ❑ *More than 10 times per week*

(Yes, I've seen stuff this bad.)

Of course, you could avoid problems with response sets by asking *open-ended questions*. They're no panacea, though. Closed- and open-ended questions have their advantages and disadvantages. Open-ended questions can give respondents freedom to answer how they choose, they remove any potential for response set bias, and they allow for rich, in-depth responses if a respondent is motivated enough. However, respondents can be shockingly ambiguous themselves, they can give responses that obviously indicate the question was misunderstood, or they can just plain answer with total nonsense. The researcher is then left with a quandary— what to do with these responses? Throw them out? Is that honest? Try to make sense of them? Is *that* honest? Closed-ended questions do have their problems, but the answers are unambiguous, and the data they generate are easy to manage. It's a tradeoff: With closed-ended questions, the researcher is structuring the data, which keeps things nice and tidy; with open-ended questions, the researcher is giving power to respondents to structure the data, which can be awfully messy, but it can also yield rich, unanticipated results.

Choosing open-ended and closed-ended questions to different degrees gives us a continuum of approaches to asking individuals questions, from loosely structured, conversational-style interviews, to highly standardized interviews, to fill-in-the-bubble questionnaires. When we conduct interviews, it is usually in a *semi-structured interview* style, with the same mostly open-ended questions asked, but with variations in wording, order, and follow-ups to make the most of the organic nature of human interaction.

When we interview a small group of people at once, it's called a *focus group*. Focus groups are not undertaken for the sake of efficiency—it's not just a way to get a lot of interviews done at once. Why do we conduct focus groups, then? When you go see a movie with a group of friends, you leave the theater with a general opinion of the movie—you liked it, you hated it, you thought it was funny, you thought it meant .... When you go out for dessert afterward and start talking with your friends about the movie, though, you find that your opinion is refined as it emerges in the course of that conversation. It's not that your opinion didn't exist before or, necessarily, that the discussion changed your opinion. Rather, it's in the course of social interaction that we uncover and use words to express our opinions, attitudes, and values that would have otherwise lain dormant. It's this kind of *emergent* opinion that we use focus groups to learn about. We gather a group of people who have something in common—a common workplace, single parenthood, Medicaid eligibility—and engage them in a guided conversation so that the researcher and participants alike can learn about their opinions, values, and attitudes.

Asking questions is central to much empirical social research, but we also collect data by directly observing the phenomena we're studying, called *field research* or simply (and more precisely, I think) *direct observation*. We can learn about political rallies by attending them, about public health departments by sitting in them, about public transportation by riding it, and about judicial confirmation hearings by watching them. In the conduct of empirical social research, such attending, sitting, riding, and watching aren't passive or unstructured. To prepare for our direct observations, we construct a *direct observation tool* (or *protocol*), which acts like a questionnaire that we "ask" of what we're observing. Classroom observation tools, for example, might prompt the researcher to record the number of students, learning materials available in the classroom, student-teacher interactions, and so on.

The advice for developing useful observation tools isn't unlike the advice for developing useful instruments for asking questions; the tool must enable an accurate, thorough, unbiased description of what's observed. Likewise, a potential pitfall of direct observation is not unlike social desirability bias: When people are being observed, their knowledge of being observed may affect their behavior in ways that bias the observations. This is the problem of *participant reactivity*. Surely the teacher subjected to the principal's surprise visit is a bit more on his game than he would have been otherwise. The problem isn't insurmountable. Reactivity usually tapers off after a while, so we can counter this problem by giving people being observed enough time to get used to it. We can just try to be unobtrusive, we can make observations as participants ourselves (*participant observation*), or, sometimes, we can keep the purpose of the study a mystery so that subjects wouldn't know how to play to our expectations even if they wanted to.

Finally, we can let other people do our data collection for us. If we're using data that were collected by someone else for their own purposes, our data collection strategy is using *secondary data*. Social science researchers are fortunate to have access to multiple online *data warehouses* that store datasets related to an incredibly broad range of social phenomena. In political science, for example, we can download and analyze general public opinion datasets, results of surveys about specific public policy issues, voting data from federal and state legislative bodies, social indicators for every country, and on and on. Popular data warehouses include Inter-University Consortium for Political and Social Research (ICPSR), University of Michigan's National Elections Studies, Roper Center for Public Opinion Research, United Nations Common Database, World Bank's World Development Indicators, and U.S. Bureau of the Census. Such secondary data sources present research opportunities that would otherwise outstrip the resources of many researchers, including students.

A particular kind of secondary data, *administrative data*, are commonly used across the social sciences, but are of special interest to those of us who do research related to public policy, public administration, and other kinds of organizational behavior. *Administrative data* are the data collected in the course of administering just about every agency, policy, and program. For public agencies, policies, and programs, they're legally accessible thanks to freedom of information statutes, and they're frequently available online. Since the 1990s, these datasets have become increasingly sophisticated due to escalating requirements for performance measurement and program evaluation. Still, beware: Administrative datasets are notoriously messy. These data usually weren't collected with researchers in mind, so the datasets require a lot of cleaning, organizing, and careful scrutiny before they can be analyzed.

Formal research designs

Simply collecting data is insufficient to answer research questions. We must have a plan, a *research design*, to enable us to draw conclusions from our observations. Different methodologists divvy up the panoply of research designs different ways; we'll use five categories: cross-sectional, longitudinal, experimental, quasi-experimental, and case study.

Cross-sectional research design is the simplest. Researchers following this design are making observations at a single point in time; they're taking a "snapshot" of whatever they're observing. Now, we can't take this too literally. A cross-sectional survey may take place over the course of several weeks. The researcher won't, however, care to distinguish between responses collected on day 1 versus day 2 versus day 28. It's all treated as having been collected in one wave of data collection. Cross-sectional research design is well suited to descriptive research, and it's commonly used to make *cross-case comparisons*, like comparing the responses of men to the responses of women or the responses of Republicans to the responses of Democrats. If we're interested in establishing causality with this research design, when we have to be sure that cause comes before effect, though, we have to be more careful. Sometimes it's not a problem. If you're interested in determining whether respondents' region of birth influences their parenting styles, you can be sure that the respondents were born wherever they were born before they developed any parenting style, so it's OK that you're asking them questions about all that at once. However, if you're interested in determining whether interest in politics influences college students' choice of major, a cross-sectional design might leave you with a chicken-and-egg problem: Which came first? A respondent's enthusiasm for following politics or taking her first political science course? Exploring causal research questions using cross-sectional design isn't verboten, then, but we do have to be cautious.

Longitudinal research design involves data collection over time, permitting us to measure change over time. If a different set of cases is observed every time, it's a *time series* research design. If the same cases are followed over time, with changes tracked at the case level, it's a *panel* design.

Experimental research design is considered by most to be the gold standard for establishing causality. (This is actually a somewhat controversial statement. We'll ignore the controversy here except to say that most who would take exception to this claim are really critical of the misapplication of this design, not the design itself. If you want to delve into the controversy, do an internet search for federally required randomized controlled trial program evaluation designs.) Let's imagine an experimental-design study of whether listening to conservative talk radio affects college students' intention to vote in an upcoming election. I could recruit a bunch of students (with whichever sampling plan I choose) and then have them all sit in a classroom listening to MP3 players through earbuds. I would have randomly given half of them MP3 players with four hours of conservative talk radio excerpts and given the other half MP3 players with four hours of muzak. Before they start listening, I'll have them respond to a questionnaire item about their likelihood of voting in the upcoming election. After the four hours of listening, I'll ask them about their likelihood of voting again. I'll compare those results, and if the talk radio group is now saying they're more likely to vote while the muzak group's intentions stayed the same, I'll be very confident in attributing that difference to the talk radio.

My talk radio experiment demonstrates the three essential features of experimental design: random assignment to experimental and control groups, control of the experimental setting, and manipulation of the independent variable. *Control* refers to the features of the research design that rule out competing explanations for the effects we observe. The most important way we achieve control is by the use of a *control group*. The students were *randomly assigned* to a control group and an *experimental group*. The experimental group gets the "treatment"—in this case, the talk radio, and the control group gets the status quo—in this case, listening to muzak. Everything else about the experimental conditions, like the time of day and the room they were sitting in, were controlled as well, meaning that the only difference in the conditions surrounding the experimental and control groups was what they listened to. This *experimental control* let me attribute the effects I observed—increases in the experimental group's intention to vote—to the cause I introduced—the talk radio.

The third essential feature of experimental design, manipulation of the independent variable, simply means the researcher determines which cases get which values of the independent variable. This is simple with MP3 players, but, as we'll see, it can be impossible with the kinds of phenomena many social researchers are interested in.

Experimental methods are such strong designs for exploring questions of cause and effect because they enable researchers to achieve the three criteria for making causal claims—the standards we use to assess the validity of causal claims: time order, association, and nonspuriousness. Time order is the easy one (unless you're aboard the starship Enterprise). We can usually establish that cause preceded effect without a problem. Association is also fairly easy. If we're working with quantitative data (as is usually the case in experimental research designs), we have a whole arsenal of statistical tools for demonstrating whether and in what way two variables are related to each other. If we're working with qualitative data, good qualitative data analysis techniques can convincingly establish association, too.

Meeting the third criterion for making causal claims, nonspuriousness, is trickier. A spurious relationship is a phony relationship. It looks like a cause-and-effect relationship, but it isn't. *Nonspuriousness*, then, requires that we establish that a cause-and-effect relationship is the real thing—that the effect is, indeed, due to the cause and not something else. Imagine conducting a survey of freshmen college students. Based on our survey, we claim that being from farther away hometowns makes students more likely to prefer early morning classes. Do we meet the first criterion? Yes, the freshmen were from close by or far away before they ever registered for classes. Do we meet the second criterion? Well, it's a hypothetical survey, so we'll say yes, in spades: Distance from home to campus and average class start time are strongly and inversely correlated.

What about nonspuriousness, though? To establish nonspuriousness, we need to think of any competing explanations for this alleged cause-and-effect relationship and rule them out. After running your ideas past the admissions office folks, you learn that incoming students from close by usually attend earlier orientation sessions, those from far away usually attend later orientation sessions, and—uh-oh—they register for classes during orientation. We now have a potential competing explanation: Maybe freshmen who registered for classes later are more likely to end up in early morning classes because classes that start later are already full. The students' registration date, then, becomes a potentially important *control variable*. It's potentially important because it's quite plausibly related to both the independent variable (distance from home to campus) and the dependent variable (average class start time). If the control variable, in fact, *is* related to both the independent variable and dependent variable, then that alone could explain why the independent and dependent variables *appear* to be related to each other when they're actually not. When we do the additional analysis of our data, we confirm that freshmen from further away did, indeed, tend to register later than freshmen from close by, that students who register later tend to end up in classes with earlier start times, and, when we control for registration date, there's not an actual relationship between distance from home and average class start time. Our initial causal claim does not achieve the standard of nonspuriousness.

The beauty of experimental design—and this is the crux of why it's the gold standard for causal research—is in its ability to establish nonspuriousness. When conducting an experiment, we don't even have to think of potential control variables that might serve as competing explanations for the causal relationship we're studying. By randomly assigning (enough) cases to experimental and control groups and then maintaining control of the experimental setting, we can assume that the two groups and their experience in the course of the study are alike in every important way except one—the value of the independent variable. Random assignment takes care of potential competing explanations we can think of *and* competing explanations that never even occur to us. In a tightly controlled experiment, any difference observed in the dependent variable at the conclusion of the experiment can confidently be attributed to the independent variable alone.

"Tightly controlled experiments," as it turns out, really aren't that common in social research, though. Too much of what we study is important only when it's out in the real world, and if you try to stuff it into the confines of a tightly controlled experiment, we're unsure if what we learn applies to the real thing. Still, experimental design is something we can aspire to, and the closer we can get to this ideal, the more confident we can be in our causal research. Whenever we have a research design that mimics experimental design but is missing any of its key features— random assignment to experimental and control groups, control of the experimental setting, and manipulation of the independent variable—we have a *quasi-experimental design*.

Often, randomly assigning cases to experimental and control groups is prohibitively difficult or downright impossible. We can't assign school children to public schools and private schools, we can't assign future criminals to zero tolerance states and more lax states, and we can't assign pregnant women to smoking and nonsmoking households. We often don't have the power to manipulate the independent variable, like deciding which states will have motor-voter laws and which won't, to test its effects on voting behaviors. Very rarely do we have the ability to control the experimental setting; even if we could randomly assign children to two different kindergarten classrooms to compare curricula, how can other factors—the teachers' personalities, for instance—truly be the same?

Quasi-experimental designs adapt to such research realities by getting as close to true experimental design as possible. There are dozens of variations on quasi-experimental design with curious names like *regression discontinuity* and *switching replications with nonequivalent groups*, but they can all be understood as creative responses to the challenge of approximating experimental design. When we divide our cases into two groups by some means other than random assignment, we don't get to use the term control group anymore, but *comparison group* instead. The closer our comparison group is to what a control group would have been, the stronger our quasi-experimental design. To construct a comparison group, we usually try to select a group of cases similar to the cases in our experimental group. So, we might compare one kindergarten classroom enjoying some pedagogical innovation to an adjacent kindergarten classroom with the same old curriculum or Alabama drivers after a new DUI law to Mississippi drivers not bound by it.

If we're comparing these two groups of drivers, we're also conducting a *natural experiment*. In a natural experiment, the researcher isn't able to manipulate values of the independent variable; we can't decide who drives in Mississippi or Alabama, and we can't decide whether or not a state would adopt a new DUI law. Instead, we take advantage of "natural" variation in the independent variable. Alabama did adopt a new DUI law, and Mississippi did not, and people were driving around in Alabama and Mississippi before and after the new law. We have the opportunity for before-and-after comparisons between two groups, it's just that we didn't introduce the variation in the independent variable ourselves; it was already out there.

Social researchers also conduct *field experiments*. In a field experiment, the researcher randomly assigns cases to experimental and comparison groups, but the experiment is carried out in a real-life setting, so experimental control is very weak. I once conducted a field experiment to evaluate the effectiveness of an afterschool program in keeping kids off drugs and such. Kids volunteered for the program (with their parents' permission). There were too many volunteers to participate all at once, so I randomly assigned half of them to participate during fall semester and half to participate during spring semester. The fall semester kids served as my experimental group and, during the fall semester, the rest of the kids served as my comparison group. At the beginning of the fall semester, I had all of them complete a questionnaire about their attitudes toward drug use, etc., then the experimental group participated in the program while the control group did whatever they normally did, and then at the end of the semester, all the kids completed a similar questionnaire again. Sure enough, the experimental group kids' attitudes changed for the better, while the comparison group kids' attitudes stayed about the same (or even changed a bit for the worse). All throughout the program, the experimental group and comparison group kids went about their lives—I certainly couldn't maintain experimental control to ensure that the only difference between the two groups was the program.

Very strong research designs can be developed by combining one of the longitudinal designs (time series or panel) with either experimental or quasi-experimental design. With such a design, we observe values of the dependent variable for both the experimental and control (or comparison) groups at multiple points in time, then we change (or observe the change of) the independent variable for the experimental group, and then we observe values of the dependent variable for both groups at multiple points in time again.

That's a bit confusing, but an example will clarify: Imagine inner-city pharmacies agree to begin stocking fresh fruits and vegetables, which people living nearby otherwise don't have easy access to. We might want to know whether this will affect area residents' eating habits. There are lots of ways we could go about this study, but probably the strongest design would be an *interrupted time series quasi-experimental design*. Here's how it might work: Before the pharmacies begin stocking fresh produce, we could conduct door-to-door surveys of people in two inner-city neighborhoods—one without a pharmacy and one with a pharmacy. We could survey households once a month for four months before the produce is stocked, asking folks about how much fresh produce they eat at home.

(A quick aside: We'd probably want to talk to different people each time since, otherwise, just the fact that we keep asking them about their eating habits, they might change what they eat—an example of a *measurement artifact*, which we try to avoid. We want to measure changes in our dependent variable, *eating habits*, that are due to change in the independent variable, *availability of produce at pharmacies*, not due to respondents' participation in the study itself.)

After the pharmacies begin stocking fresh produce, we would then conduct our door-to-door surveys in both neighborhoods again, perhaps repeating them once a month for another four months. Once we're done, we'd have a very rich dataset for estimating the effect of available produce on eating habits. We could compare the two neighborhoods before the produce was available to establish just how similar their eating habits were before, and then we could compare the two neighborhoods afterward. We might see little difference one month after the produce became available as people became aware of it, then maybe a big difference in the second month in response to the novelty of having produce easily available, and then maybe a more moderate, steady difference in the third and fourth months as some people returned to their old eating habits and others continued to purchase the produce. With this design, we can provide very persuasive evidence that the experimental and comparison groups were initially about the same in terms of the dependent variable, which increases our confidence that any changes we see later are indeed due to the change in the independent variable. We can also capture change over time, which is frequently very important when we're measuring behavioral changes, which tend to diminish over time.

Case study research design is the oddball of the formal research designs. Many researchers who feel comfortable with all the other designs would feel ill equipped to undertake a case study. A *case study* is the systematic study of a complex case that is in-depth and holistic. Unlike the other designs, we're just studying a single case, which is usually something like an event, such as a presidential election, or a program, such as the operation of a needle exchange program. With the other designs, we usually rely on a single data collection method, but with case study research design, we use multiple data collection methods, with a heavy emphasis on collecting qualitative data. In the course of a single case study, we might conduct interviews, conduct focus groups, administer questionnaires, survey administrative records, and conduct extensive direct observations. We make enough observations in as many different ways as necessary to enable us to write a rich, detailed description of our case. This written report is, itself, called a *case study*.

The richness of case studies highlights another key difference between this and the other research designs. The contrast with experimental design is sharpest: If you think about experimental design, its beauty lies in *ignoring* complexity. If I were to randomly assign a bunch of teenagers to experimental and control groups, my express intention would be to ignore all their pimply, hormonal, awkward, exuberant complexity and the group dynamics that would undoubtedly emerge in the two groups. I count on random assignment and experimental control to make all differences between the two groups a complete wash except the difference in the independent variable. With case studies, though, we embrace this complexity. The whole point is to describe this rich complexity, bringing only enough organization to it to make it understandable to people who can't observe it directly—those people who will ultimately read our written case studies.

There are many elaborations on these formal research designs. A few more, along with a system of notation for depicting research designs, are presented in Appendix B.

DATA ANALYSIS

"Let the data speak for itself" is a frequently invoked dictum that is both grammatically incorrect and impossible. Data, having been recorded, do not then speak for themselves. Data have no meaning apart from how we interpret them. *Data analysis* is the task of finding meaningful patterns in our data. It's how we make sense of our data, how we derive meaning from it.

It is accurate enough to say that *quantitative data analysis* helps us make sense of numeric data and *qualitative data analysis* helps us make sense of textual data, but that does oversimplify the distinction a bit. Imagine conducting direct observations of presidential primary campaign stump speeches. Each time we observe a speech, we would probably want to record the approximate number of people in attendance. Clearly, that will yield numeric data, and we would use quantitative data analysis techniques to find patterns in them, such as calculating the mean, median, and standard deviation to summarize the central tendency and variation of crowd sizes at the speeches. We would probably also record the speeches themselves and later transcribe them so that we have a verbatim written record of each speech. This time, we will, clearly, have textual data and use qualitative data analysis tools to identify underlying themes in the data. However, we would also record whether each speech was delivered by a Republican primary candidate or a Democrat primary candidate, probably by checking a box on our direct observation tool. In this case, the data we record is, in a sense, qualitative; it's text, *Republican* or *Democrat*. When we analyze these data, though, we will most likely use quantitative data analysis tools, in this case, probably just to count the frequency of each value of the variable, political party. The choice between qualitative and quantitative data analysis tools, then, isn't entirely about the type of data; it's also determined by what we're going to do with those data. If we're performing numeric calculations, we use quantitative data analysis tools, and if we're deriving and attributing meaning from and to words, we use qualitative data analysis tools. (Even that oversimplifies a little because of gray areas like *content analysis*, which is a quantitative approach to qualitative data analysis, but we'll leave it there.)

The processes of qualitative data analysis and quantitative data analysis differ as well. When we undertake quantitative data analysis, the concepts we're measuring are almost always predetermined. We first decide to measure a concept like political literacy, then operationalize the concept by writing a list of quiz items, then collect our data, and, finally, tally our respondents' scores—that is, conduct our quantitative data analysis—as an indicator of their political literacy. Conceptualization came first, analysis second. When we're doing qualitative data analysis, though, this isn't necessarily the case. If we want to conduct interviews to understand (in the *verstehen* sense, recall) what respondents believe it means to be politically literate, we may not know what concepts we'll end up identifying—that's why we're doing the research. Certainly, we have some starting point—a formal theory, a model, a hunch, whatever we've learned from previous research—or we wouldn't know what to ask questions about. It is during the course of data analysis, though, that we identify important concepts as we find patterns in our interview data. Thus, conceptualization and analysis are pursued iteratively; concepts are a starting point for data collection, consistent with our model of the research process, but concepts are also the *product* of qualitative data analysis.

Much more of the quantitative data analysis process is a settled matter than the qualitative data analysis process. There is only one way to calculate the sample standard deviation, and if you want to compare the means of two groups, there are nearly universally agreed upon rules to help you choose the appropriate statistical test. If you want to identify underlying themes in a political speech, though, there is not one right way to go about your analysis. There are many different qualitative data analysis camps, some complementary and some competing, and even within one camp, there is no expectation that qualitative data analysis would lead you and another researcher to precisely the same findings.

We're not going to cover the "how to" of data analysis here. For that, I refer you to your introductory statistics and qualitative data analysis courses and textbooks. Most students reading this will also have an introductory statistics course. I think we do aspiring social science researchers a disservice by not also requiring a course in qualitative data analysis. Students find one final distinction appealing. The frank truth is that students can accomplish little high caliber research, by professional standards, using the quantitative data analysis tools learned in an introductory statistics course. There are exceptions, but the type of quantitative research that could be published in a social science journal generally requires more statistics training. In contrast, students can conduct excellent research using basic qualitative data analysis techniques—a lot of good work is done with the basic tools. You shouldn't choose your data analysis methods based on this, of course, but you should be encouraged to know that qualitative data analysis skills are accessible and can enable students to conduct strong research. Great starting points are David Thomas's (2006) "A General Inductive Approach for Analyzing Qualitative Evaluation Data," *American Journal of Evaluation, 27*(2), 237-246; and Virginia Braun and Victoria Clarke's (2021) *Thematic Analysis: A Practical Guide* (Sage).

I find that students often show up in my research methods courses still just a little uncertain about inferential statistics, even if they're fresh out of a statistics course. That's not a criticism of the students or their statistics courses (sometimes it's my own course!)—it's a hard idea to grasp at first. If you're one of those uncertain students, I offer a quick review of this data analysis approach in Appendix C.

One final note about data analysis: Incorporating control variables into data analysis often trips students up. Appendix D presents one way of approaching this called *elaboration modeling*. I like to introduce students to this strategy because its logic can be applied across a wide range of quantitative and qualitative data analysis scenarios, and it helps students better learn the concept of control as well.

GENERALIZING AND THEORIZING

When we've completed our data analysis, it's time to complete the loop, which entails at least three somewhat overlapping tasks. First, we return to our research question, not asking it, but answering it. What did we learn? To what extent can we generalize our findings—to a larger population, to other settings, to other cases, to other times? We do this humbly. In social research, the claims we make are almost always provisional. We rarely claim to wholly "answer" a research question, and we virtually never claim to "prove" anything. We state our conclusions tentatively, realizing that future research could improve on, expand, or even contradict what we've learned. We also acknowledge the limitations of our own research (which are always present) and suggest directions for future research. (See Appendix F about how to avoid the common error of generalizing from groups to individuals.) Second, we relate what we've learned to previous research. How is what we've learned consistent with previous research? How is it different? Where does it fit in to the larger body of knowledge? Third, we advance what we know about theory. From a deductive perspective, does the theory that drove our research seem to be a good fit with what we've observed? How might our observations suggest we should modify the theory? From an inductive perspective, what theory did we construct based on our observations? How might future research test this theory?

EVALUATING RESEARCH: VALIDITY AND RELIABILITY

As you may have surmised, doing research is not exactly a science. You may have noticed that I switch between "social science research" and "social research." I'm ambivalent on whether what we do is "science," exactly—it depends what you mean by "science," and smart people disagree on that point. I'm at peace with my ambivalence. While writing, I've been self-conscious about how I'm constantly qualifying my statements—I've used the word *usually* 37 times so far, and *sometimes*, 30. That's not the mark of particularly good writing, but it does reflect an important point: There is not one right way to do any research project. When we're making decisions about how to go about our research, we're faced with many options. Identifying these options is a creative process; we brainstorm, we trade ideas with others, we tease out the implications of our theoretic bases, we look to previous research for inspiration, and we're left with a myriad of options. If we're interested in learning about public managers' leadership styles, we could interview them, conduct focus groups with them, have them complete a web survey, or observe them in action. We could structure our observations in a cross-sectional research design, make cross-case comparisons, follow managers over time, or devise a clever experiment. When it comes to operationalizing any one of the many concepts we need to measure, we're faced with still more choices. To decide how to operationalize a concept like *transformational leadership*, we'll look to our fellow researchers, theories, and previous research, but we'll still be left with infinite variations on how we could ask questions, extract data from administrative records, or record direct observations.

As creative as doing research is, however, it would be misleading to say that doing research is an art. It is a creative endeavor to be sure, but it's definitely not the case that what constitutes good research is "in the eye of the beholder." It's more like a craft. Doing research takes a lot of creativity, but it can be done well or poorly. Doing research is not a wholly subjective enterprise; there are standards that we can apply to judge research quality. Broadly speaking, the two standards used to judge the quality of research are *validity* and *reliability*. We use these terms as special bits of jargon in research methodology, where they take on meaning beyond what we mean when using them colloquially. (And to pile the po-mo even higher, I should note that of all the jargon we've covered, the jargon related to validity and reliability is the most inconsistently applied among social science methodologists. Methodologists all seem to have their own twist on how they use these terms, so understand that you're about to get my distillation of all that, and it won't necessarily always jibe with how you'll see the terms used elsewhere.) We should know how to apply these standards because it helps us decide how much stock to put in research that we read and because knowing the standards by which research is judged helps us design research ourselves that will meet those standards.

We can think of evaluating research design on two levels: overall research design and operationalization of specific concepts. For any given research project, then, we can make holistic evaluations of the merits of the entire project, and we can also make evaluations of how each individual concept was measured, which could amount to dozens of discrete evaluations for a single research project.

When we're evaluating the overall design of a research project, we apply the standards of internal validity, external validity, and reliability. *Internal validity* is the extent to which the inferences we make from our observations are true. Most often, the standard of internal validity is applied to causal inferences. If we assess a study's internal validity, then, we're assessing the degree to which the design of that study permits confident inferences about cause and effect. Experimental designs, when well done, are very high in internal validity; we can be confident that the observed changes in the dependent variable are, indeed, due to the changes in the independent variable. It's important to see that strong internal validity is a function of the research design; characteristics of the research design itself—in the case of experiments, the random assignment of cases to experimental and control groups and the control of the experimental setting—allow us to make our causal claims with a lot of confidence.

Interestingly enough, the characteristics of experiments that strengthen internal validity are the same characteristics that tend to weaken external validity. *External validity* is the extent to which we can generalize the inferences we make from observations beyond the cases observed. Assessing external validity asks whether or not we can apply what we've learned from our observations to other cases, settings, or times. When we conduct an experiment, it's usually very artificial—the whole setting of the experiment has to be tightly controlled to ensure comparability of the experimental and control groups in every respect except their values for the independent variable. (I hope you thought about that when you read about students listening to conservative talk radio through their earbuds for four hours straight while sitting in a classroom—not a very realistic scenario.) This tight control is essential to achieving internal validity, but it makes it really hard to apply it to other settings (like real life)—it makes it hard to achieve external validity.

Reliability is the extent to which other researchers would get the same results if the study were repeated, whether by themselves or by someone else. Most often, assessing reliability is a thought experiment—an exercise we carry out only in our imaginations. Let's return to the example of surveying people in inner-city neighborhoods about their eating habits. If I were to assess the reliability of our quasi-experimental research design, I would think through a few hypothetical scenarios. What if someone else had conducted this study? I'm a white male; what if a black female had conducted the interviews instead? Would she have gotten the same results as me? What if I could hit the cosmic reset button, go back in time, and conduct the study again myself? Would I, myself, get the same results again?

When we evaluate a study at the level of the operationalization of all its concepts, we apply the standards of operational validity and, again, reliability. *Operational validity* is the extent to which the way we have operationalized a concept truly measures that concept. Let's consider the challenge of operationalizing a concept college students are familiar with, *college readiness*. If I were to take a stab at a nominal definition of *college readiness*, I'd say something like "a person's preparedness for success in college." How might we operationalize this concept? We have lots of options, but let's say we're going to administer a written questionnaire to college applicants, and we'll include the following question as our measure of college readiness:

> *What was your score on the ACT?*

That seems straightforward enough, but let's evaluate this operationalization of college readiness in terms of its operational validity. Does this question *really* measure college readiness? We can assess operational validity from four different angles: face validity, content validity, discriminate validity, and criterion validity. (In introducing these terms, I should mention a quibble I have with lots of textbook authors. These aren't really different *types* of validity; they're all different aspects of operational validity—different ways of thinking about whether or not an operationalization really measures the concept it's intended to measure.)

Face validity is the most intuitive of these four ways to think about operational validity. When we assess the face validity of an operationalization, we're just asking whether, on the face of it, the operationalization seems to measure its targeted concept. Here, I'd say sure—it seems very reasonable to use ACT scores as a measure of college readiness. As evidence for the face validity of this operationalization, I could refer to other researchers who have used this same operationalization to measure college readiness. Certainly, ACT score achieves face validity as a measure of college readiness.

Next, we can think about operational validity by assessing the measure's *content validity* (sometimes called *construct validity*). Many abstract concepts we want to measure are broad and complex. Think about college readiness. Surely it includes academic readiness, which itself is multifaceted—having adequate studying skills, critical thinking skills, math skills, writing skills, computer skills, and so on. College readiness probably also includes nonacademic factors as well, like self-motivation, openness to new ideas, ability to get along well in a group, and curiosity. I'm sure you can think of still more aspects of college readiness. When we assess content validity, we ask whether or not our operationalization measures the full breadth and complexity of a concept. Here, I think our ACT score might be in trouble. Of all the many aspects of college readiness, ACT scores only measure a swath of the academic skills. Those academic skills are, indeed, indicators of college readiness (and hence ACT scores do achieve face validity), but if we're relying solely on ACT scores as our full operationalization of college readiness, our operationalization fails to achieve content validity. We almost always require multiple measures when operationalizing complex concepts in order to achieve content validity.

At this point in our research design, we'd probably add some additional items to our questionnaire to operationalize college readiness more fully. Let's continue, though, assessing our original operationalization, relying only on ACT scores as a measure of college readiness. We can continue to assess the operational validity of this operationalization by assessing its *discriminate validity*, which asks whether or not the way we've operationalized our concept will enable us to distinguish between the targeted concept and other concepts. We all had a friend in high school who didn't do so hot on the ACT and unwittingly attributed the poor showing to discriminate validity: "ACT scores just show how good you are at taking standardized tests!" Your friend was saying that the ACT doesn't operationalize the concept it's intended to operationalize, college readiness, but another concept altogether, standardized-test-taking ability. Your friend was quite astute to consider whether the ACT achieves discriminate validity.

If considering face validity is the most intuitive way of assessing operational validity, considering *criterion validity* is the most formal. When we assess criterion validity, we test, usually statistically, whether or not our measures relate to other variables as they should if we have successfully operationalized our target concept. If ACT score successfully operationalizes college readiness, what should students' ACT scores be statistically associated with? Well, if ACT scores really are a measure of college readiness, then students who had higher ACT scores should also tend to have higher college GPAs. If we test for that association, we're using college GPA as a *criterion variable* (hence *criterion* validity) for determining whether or not ACT scores are a good way to operationalize college readiness. If there's a strong association between ACT scores (the variable we're testing) and college GPA (our criterion variable), then we'll use that as evidence that our operationalization of college readiness (our target concept) demonstrates operational validity. We could think of other criterion variables as well—whether or not the student graduates from college and how long it takes come to mind. We don't always have the opportunity to test for criterion validity, but when we do, it can provide very strong evidence for our measures' operational validity.

Just as when we were evaluating the overall research design, we apply the standard of reliability when we evaluate the operationalization of an individual concept, likewise engaging in thought experiments to consider whether we'd get the same results if the observations were made by other researchers or even by ourselves if we could go back and do it again. We also consider, and sometimes quantify using statistical tools, the degree to which individual measures demonstrate *random error*. This is the amount of variation in repeated measures, whether repeated in reality or only hypothetically. Say we're measuring the height of a wall using a tape measure. We know that the wall's height is 96 inches. You can imagine, though, that your tape measure might read 95 ⅞ the first time you measure it, 96 ⅛ the second time, and 95 $^{15}/_{16}$ the third time. Your measurement is exhibiting some random error. If you were to repeat this over and over, the mean measurement would be about right, but any one measurement is bound to be off just a little.

In social research, some types of measures are more susceptible to random error than others. Imagine being asked to rate your agreement or disagreement with the statement *I like campaign signs printed in all caps* on a 7-point scale. I know I don't have a particularly strong opinion on the matter, really. If you asked me this morning, I might rate it a 5, but this afternoon, it might be a 4, and tomorrow it might be a 7. We very rarely actually take measurements from the same cases over and over again (and if you did, I'd probably start always giving you the same answer anyway just for the sake of sounding consistent with myself), so we have to think about the consistency of hypothetical repeated measurements. Hypothetically, if we were to ask someone to rate how much he likes campaign signs in all caps, zap his memory of the experience, ask him again, zap, ask again, zap, ask again, zap, and ask again, I'd predict that we'd observe a lot of random error, meaning our question is probably not a very reliable way to operationalize the targeted concept, preference for capitalization of campaign sign text.

RESEARCH ETHICS

When studying human behavior, opportunities for unethical behavior abound. Human nature being what it is, researchers must be on their guard against unethical research practices. There's a lot of temptation to lie. If you want to make a big name for yourself as a researcher, or if you're hoping to use research to support your opinion, it's tempting to fabricate data or falsify findings to suit your needs, especially when the actual findings are a dud. We've seen that we incorporate what we learn from previous research throughout the research process, and when doing so, we are always careful to cite sources of words and ideas that are not our own.

When we are collecting data from people—interviewing them, observing them, rifling through their administrative records—we make every effort not to harm them. We make sure research participants know of any potential risks of participating in our studies, including obvious things like physical harm, of course, but also including the risk that their personal information— however unrisky we may think this is—will become known to others. Often, we promise our research participants confidentiality, and we work hard to meet that ethical commitment. Our research participants are not merely "subjects," they are neither data points nor ID numbers, they cannot be fully known by the values we assign to variables for them, and they are not individual representatives of the generalizations we hope to derive from our research (see Appendix F on this last point). The people who participate in research are individuals of inestimable worth and dignity, and they should be respected accordingly.

When we conduct research under the auspices of a university or government agency, our research ethics are monitored by people appointed to their Institutional Review Boards (IRBs). IRBs certify that researchers have been trained in research ethics, usually by verifying that researchers have completed an online training module. We submit our research plans to these boards, including plans for how we will ensure the ethicality of our research projects, and we wait for the green light from them before we proceed. They monitor our progress and serve as a point of contact for anyone needing to express a concern about the ethical conduct of researchers. To be honest, IRBs can be a bit of a hassle to the researcher just wanting to get on with the fun work of doing research, but their responsibilities, particularly the protection of human research subjects, are indispensable to ensuring the ethicality of social research.

CONCLUDING REMARKS

I hope you now agree that how to do empirical social science research is not a mystery, and learning from and evaluating others' research is something you, yourself, can do. To keep learning, read reports of previous research—lots of them. No matter what sorts of social phenomena you're interested in, there is a body of research about it, and it's now more accessible to you than ever. A few of you will conduct research yourself, and if you're interested in doing research as a career, you should get as much practical research experience as you can, starting now. *All* of you can use what you've learned here by consuming research as engaged citizens who can make sense of and participate in the empirical arguments that enter public discourse in our workplaces, communities, states, nation, and the world. Please do.

DATA COLLECTION METHODS

The decision of how to select cases to observe may present a long list of options, but deciding what specific types of data to collect presents us with infinite options. It seems to me, though, that the kinds of data collection we do in empirical social research all fall in one of three broad categories: asking questions, making direct observations, and collecting secondary data.

Collecting data by asking questions can be somewhat like our everyday experience of carrying on conversations. If you have taken an introductory communications course, you have learned how interpersonal communication involves encoding our intended meaning in words, transmitting those words to our conversation partner, who then receives those words, decodes them to derive meaning, and then repeats the process in response. All of this can be derailed due to distractions, assumptions, moods, attitudes, social pressures, and motives. In normal conversation, both parties can try to keep communication on track by reading body language, asking clarifying questions, and correcting misunderstandings. When asking questions for research, though, you—the researcher—are solely responsible for crafting a question-and- answer exchange that yields valid data. The researcher must ensure the meaning she intends to encode in her questions are accurately decoded by the respondent; she must ensure the respondent is enabled to accurately encode his intended meaning in his available response options; she must anticipate and mitigate threats to the accurate encoding and decoding of meaning posed by those distractions, assumptions, moods, attitudes, social pressures, and motives. Before thinking about the nuts and bolts of asking questions for research, understand that it is, essentially, two-way communication with all responsibility for ensuring its accuracy on the head of the researcher.

Volumes have been written about the craft of asking people questions for research purposes, but we can sum up the main points briefly. Researchers ask people questions face-to-face (whether in person or via web-based video conferencing), by telephone, using self-administered written questionnaires, and in web-based surveys. Each of these *modes of administration* has its advantages and disadvantages. It's tempting to think that face-to-face interviewing is always the best option, and often, it *is* a good option. Talking to respondents face-to-face makes it hard for them to stop midway through the interview, gives them the chance to ask questions if something needs clarifying, and lets you read their body language and facial expressions so you can help if they look confused. A face-to-face interview gives you a chance to build rapport with respondents, so they're more likely to give good, thorough answers because they want to help you out. That's a double-edged sword, though: Having you staring a respondent in the face might tempt him to give answers that he thinks you want to hear or that make him seem like a nice, smart, witty guy—the problem of *social desirability bias*.

Combating bias is one of the most important tasks when designing a research project. *Bias* is any systematic distortion of findings due to the way that the research is conducted, and it takes many forms. Imagine interviewing strangers about their opinions of a particular political candidate. How might their answers be different if the candidate is African-American and the interviewer is white? What if the respondent is interviewed at her huge fancy house and the interviewer is wearing tattered shoes? The human tendencies to want to be liked, to just get along, and to avoid embarrassment are very strong, and they can strongly affect how people answer questions asked by strangers. To the extent that respondents are affected similarly from interview to interview, the way the research is being conducted has introduced bias.

So, then, asking questions face-to-face may be a good option sometimes, but it may be the inferior option if social desirability bias is a potential problem. In those situations, maybe having respondents answer questions using a self-administered written questionnaire would be better. Completing a questionnaire in private goes a long way in avoiding social desirability bias, but it introduces other problems. Mail is easier to ignore than someone knocking at your door or making an appointment to meet with you in your office. You have to count more on the respondent's own motivation to complete the questionnaire, and if motivated respondents' answers are systematically different than unmotivated nonrespondents, your research plan has introduced *self-selection bias*. You're not there to answer questions the respondent may have, which pretty much rules out complicated questionnaire design (such as questionnaires with a lot of skip patterns—"If 'Yes,' go to Question 38; if 'No,' go to Question 40" kind of stuff). On the plus side, it's much easier and cheaper to mail questionnaires to every state's director of human services than to visit them all in person.

You can think through how these various pluses and minuses would play out with surveys administered by telephone. If you're trying to talk to a representative sample of the population, though, telephone surveys have another problem. Think about everyone you know under the age of 30. How many of them have telephones—actual land lines? How many of their parents have land lines? Most telephone polling is limited to calling land lines, so you can imagine how that could introduce *sampling bias*—bias introduced when some members of the population are more likely to be included in a study than others. When cell phones are included, you can imagine that there are systematic differences between people who are likely to answer the call and those who are likely to ignore the unfamiliar Caller ID—another source of sampling bias. If you are a counseling center administrator calling all of your clients, this may not be a problem; if you are calling a randomly selected sample of the general population, the bias could be severe.

Web-based surveys have become a very appealing option for researchers. They are incredibly cheap, allow complex skip patterns to be carried out unbeknownst to respondents, face no geographic boundaries, and automate many otherwise tedious and error-prone data entry tasks. For some populations, this is a great option. I once conducted a survey of other professors, a population with nearly universal internet access. For other populations, though—low-income persons, homeless persons, disabled persons, the elderly, and young children—web-based surveys are often unrealistic.

Deciding what medium to use when asking questions is probably easier than deciding what wording to use. Crafting useful questions and combining them into a useful data collection instrument take time and attention to details easily overlooked by novice researchers. Sadly, plentiful examples of truly horribly designed surveys are easy to come by. Well-crafted questions elicit unbiased responses that are useful for answering research questions; poorly crafted questions do not.

So, what can we do to make sure we're asking useful questions? There are many good textbooks and manuals devoted to just this topic, and you should definitely consult one if you're going to tackle this kind of research project yourself. Tips for designing good data collection instruments for asking questions, whether questionnaires, web-based surveys, interview schedules, or focus group protocols, boil down to a few basics.

Perhaps most important is paying careful attention to the wording of the questions themselves. Let's assume that respondents want to give us accurate, honest answers. For them to do this, we need to word questions so that respondents will interpret them in the way we want them to, so we have to avoid ambiguous language. (What does *often* mean? What is *sometimes*?) If we're providing the answer choices for them, we also have to provide a way for respondents to answer accurately and honestly. I bet you've taken a survey and gotten frustrated that you couldn't answer the way you wanted to.

I was once asked to take a survey about teaching online. One of the questions went something like this:

Do you think teaching online is as good as teaching face-to-face?

- ❑ *Yes*
- ❑ *No*
- ❑ *I think they're about the same*

I've taught online lot, I've read a lot about online pedagogy, I've participated in training about teaching online, and this was a frustrating question for me. Why? Well, if I answer *no*, my guess is that the researchers would infer that I think online teaching is inferior to face-to-face teaching. What if I am an online teaching zealot? By *no*, I may mean that I think online teaching is superior to face-to-face! There's a huge potential for disconnect between the meaning the respondent attaches to this answer and the meaning the researcher attaches to it. That's my main problem with this question, but it's not the only one. What is meant, exactly, by *as good as*? *As good as* in terms of what? In terms of student learning? For transmitting knowledge? My own convenience? My students' convenience? A respondent could attach any of these meanings to that phrase, regardless of what the researcher has in mind. Even if I ignore this, I don't have the option of giving the answer I want to—the answer that most accurately represents my opinion—*it depends*. What conclusions could the researcher draw from responses to this question? Not much, but uncritical researchers would probably report the results as filtered through their own preconceptions about the meanings of the question and answer wording, introducing a pernicious sort of bias—difficult to detect, particularly if you're just casually reading a report based on this study, and distorting the findings so much as to actually convey the opposite of what respondents intended. (I was so frustrated by this question and fearful of the misguided decisions that could be based on it that I contacted the researcher, who agreed and graciously issued a revised survey—research methods saves the day!) Question wording must facilitate unambiguous, fully accurate communication between the researcher and respondent.

Just as with mode of administration, question wording can also introduce social desirability bias. Leading questions are the most obvious culprit. A question like *Don't you think public school teachers are underpaid?* makes you almost fall over yourself to say "Yes!" A less leading question would be *Do you think public school teachers are paid too much, paid too little, or paid about the right amount?* To the ear of someone who doesn't want to give a bad impression by saying the "wrong" answer, all of the answers sound acceptable. If we're particularly worried about potential social desirability bias, we can use *normalizing statements*: *Some people like to follow politics closely and others aren't as interested in politics. How closely do you like to follow politics?* would probably get fewer trying-to-sound-like-a-good-citizen responses than *Do you stay well informed about politics?*

Closed-ended questions—questions that give answers for respondents to select from—are susceptible to another form of bias, *response set bias*. When respondents look at a range of choices, there's subconscious pressure to select the "normal" response. Imagine if I were to survey my students, asking them:

How many hours per week do you study?

❑ *Less than 10*
❑ *10 – 20*
❑ *More than 20*

That middle category just looks like it's the "normal" answer, doesn't it? The respondent's subconscious whispers "Lazy students must study less than 10 hours per week; more than 20 must be excessive." This pressure is hard to avoid completely, but we can minimize the bias by anticipating this problem and constructing response sets that represent a reasonable distribution.

Response sets must be exhaustive—be sure you offer the full range of possible answers—and the responses must be mutually exclusive. How *not* to write a response set:

How often do you use public transportation?

❑ *Never*
❑ *Every day*
❑ *Several times per week*
❑ *5 – 6 times per week*
❑ *More than 10 times per week*

(Yes, I've seen stuff this bad.)

Of course, you could avoid problems with response sets by asking *open-ended questions*. They're no panacea, though. Closed- and open-ended questions have their advantages and disadvantages. Open-ended questions can give respondents freedom to answer how they choose, they remove any potential for response set bias, and they allow for rich, in-depth responses if a respondent is motivated enough. However, respondents can be shockingly ambiguous themselves, they can give responses that obviously indicate the question was misunderstood, or they can just plain answer with total nonsense. The researcher is then left with a quandary— what to do with these responses? Throw them out? Is that honest? Try to make sense of them? Is *that* honest? Closed-ended questions do have their problems, but the answers are unambiguous, and the data they generate are easy to manage. It's a tradeoff: With closed-ended questions, the researcher is structuring the data, which keeps things nice and tidy; with open-ended questions, the researcher is giving power to respondents to structure the data, which can be awfully messy, but it can also yield rich, unanticipated results.

Choosing open-ended and closed-ended questions to different degrees gives us a continuum of approaches to asking individuals questions, from loosely structured, conversational-style interviews, to highly standardized interviews, to fill-in-the-bubble questionnaires. When we conduct interviews, it is usually in a *semi-structured interview* style, with the same mostly open-ended questions asked, but with variations in wording, order, and follow-ups to make the most of the organic nature of human interaction.

When we interview a small group of people at once, it's called a *focus group*. Focus groups are not undertaken for the sake of efficiency—it's not just a way to get a lot of interviews done at once. Why do we conduct focus groups, then? When you go see a movie with a group of friends, you leave the theater with a general opinion of the movie—you liked it, you hated it, you thought it was funny, you thought it meant .... When you go out for dessert afterward and start talking with your friends about the movie, though, you find that your opinion is refined as it emerges in the course of that conversation. It's not that your opinion didn't exist before or, necessarily, that the discussion changed your opinion. Rather, it's in the course of social interaction that we uncover and use words to express our opinions, attitudes, and values that would have otherwise lain dormant. It's this kind of *emergent* opinion that we use focus groups to learn about. We gather a group of people who have something in common—a common workplace, single parenthood, Medicaid eligibility—and engage them in a guided conversation so that the researcher and participants alike can learn about their opinions, values, and attitudes.

Asking questions is central to much empirical social research, but we also collect data by directly observing the phenomena we're studying, called *field research* or simply (and more precisely, I think) *direct observation*. We can learn about political rallies by attending them, about public health departments by sitting in them, about public transportation by riding it, and about judicial confirmation hearings by watching them. In the conduct of empirical social research, such attending, sitting, riding, and watching aren't passive or unstructured. To prepare for our direct observations, we construct a *direct observation tool* (or *protocol*), which acts like a questionnaire that we "ask" of what we're observing. Classroom observation tools, for example, might prompt the researcher to record the number of students, learning materials available in the classroom, student-teacher interactions, and so on.

The advice for developing useful observation tools isn't unlike the advice for developing useful instruments for asking questions; the tool must enable an accurate, thorough, unbiased description of what's observed. Likewise, a potential pitfall of direct observation is not unlike social desirability bias: When people are being observed, their knowledge of being observed may affect their behavior in ways that bias the observations. This is the problem of *participant reactivity*. Surely the teacher subjected to the principal's surprise visit is a bit more on his game than he would have been otherwise. The problem isn't insurmountable. Reactivity usually tapers off after a while, so we can counter this problem by giving people being observed enough time to get used to it. We can just try to be unobtrusive, we can make observations as participants ourselves (*participant observation*), or, sometimes, we can keep the purpose of the study a mystery so that subjects wouldn't know how to play to our expectations even if they wanted to.

Finally, we can let other people do our data collection for us. If we're using data that were collected by someone else for their own purposes, our data collection strategy is using *secondary data*. Social science researchers are fortunate to have access to multiple online *data warehouses* that store datasets related to an incredibly broad range of social phenomena. In political science, for example, we can download and analyze general public opinion datasets, results of surveys about specific public policy issues, voting data from federal and state legislative bodies, social indicators for every country, and on and on. Popular data warehouses include Inter-University Consortium for Political and Social Research (ICPSR), University of Michigan's National Elections Studies, Roper Center for Public Opinion Research, United Nations Common Database, World Bank's World Development Indicators, and U.S. Bureau of the Census. Such secondary data sources present research opportunities that would otherwise outstrip the resources of many researchers, including students.

A particular kind of secondary data, *administrative data*, are commonly used across the social sciences, but are of special interest to those of us who do research related to public policy, public administration, and other kinds of organizational behavior. *Administrative data* are the data collected in the course of administering just about every agency, policy, and program. For public agencies, policies, and programs, they're legally accessible thanks to freedom of information statutes, and they're frequently available online. Since the 1990s, these datasets have become increasingly sophisticated due to escalating requirements for performance measurement and program evaluation. Still, beware: Administrative datasets are notoriously messy. These data usually weren't collected with researchers in mind, so the datasets require a lot of cleaning, organizing, and careful scrutiny before they can be analyzed.

SAMPLING

The selection of cases to observe is the task of *sampling*. If you're going to be collecting data from people, you might be able to talk to every person that you want your research to apply to, that is, your *population*. If you're doing a study of state election commissioners, you might be able to talk to all 50 of them. In that case, you'd be conducting a *census* study. Often, though, we're only able to collect data from a portion of the population, or a *sample*. We devise a *sampling frame*, a list of cases we select our sample from—ideally, a list of all cases in the population—but then which cases do we select for the sample? We select cases for our sample by following a *sampling design*, which comes in two basic varieties: probability sampling designs and nonprobability sampling designs.

In *probability sampling designs*, every case in the population has a known, greater-than-zero probability of being selected for the sample. This feature of probability sampling designs, along with the wonder of the central limit theorem and law of large numbers, allows us to do something incredibly powerful. If we're collecting quantitative data from our sample, we can use these data to calculate *statistics*—quantified summaries of characteristics of the sample, like the median of a variable or the correlation between two variables. If we've followed a probability sampling design, we can then use statistics to estimate the *parameters*—the corresponding quantified characteristics of the population—with known levels of confidence and accuracy. This is what's going on when you read survey results in the newspaper: "± 3 points at 95% confidence." For example, if 30% of people in our sample say they'd like to work for government, then we'd be confident that if we were to repeat this survey a thousand times, 95% of the time (our level of confidence), we'd find that between 27 and 33% (because ± 3 points is our degree of accuracy) of the respondents would answer the same way. Put another way, we'd be 95% certain that 27 to 33% of the population would like to work for government.

Again, this trick of using sample statistics to estimate population parameters with known levels of confidence and accuracy only works when we've followed a probability sampling design. The most basic kind of probability sampling design is a *simple random sample*. In this design, each case in the population has a known and *equal* probability of being selected for the sample. When social researchers use the term *random*, we don't mean haphazard. (This word has become corrupted since I was in college, when my future sister-in-law started saying stuff like "A boy I knew in kindergarten just called—that was so random!" and "I just saw that guy from 'Saved by the Bell' at the mall—pretty random!") It takes a plan to be random, to give every case in the population an equal chance of being selected for a sample. If we were going to randomly select 20 state capitals, we wouldn't just select the first 20 working from west to east or the first 20 we could think of—that would introduce *sampling bias*. (We'll have more to say about *bias* later, but you get the gist of it for now.) To ensure all 50 capitals had an equal probability of being selected (a probability of 0.4, in fact), we could list them all out on a spreadsheet, use a random number generator to assign them all random numbers, sort them by those numbers, and select the first 20; or we could write each capital's name on same-sized pieces of paper, put them in a bag, shake them up, and pull out 20 names. (Some textbooks still have random number tables in the back, which you're welcome to learn how to use on your own, but they've become pretty obsolete.)

Selecting a simple random sample may be too much of a hassle because you just have a long, written list in front of you as your sampling frame, like a printed phonebook. Or, selecting a simple random sample may be impossible because you're selecting from a hypothetically infinite number of cases, like the vehicles going through an intersection. In such scenarios, you can approximate a random sample by selecting every 10^{th} or 20^{th} or 200^{th} or $whatever^{th}$ case to reach your desired sample size, which is called *systematic sampling*. This works fine as long as *periodicity* isn't present in your population, meaning that there's nothing odd about every 10^{th} (or $whatever^{th}$) case. If you were sampling evenings to observe college life, you wouldn't want to select every 7^{th} case, or you'd introduce severe sampling bias. Just imagine trying to describe campus nightlife by observing only Sunday evenings or only Thursday evenings. As long as periodicity isn't a problem, though, systematic sampling approximates simple random sampling.

Our goal in selecting a random (or systematic) sample is to construct a sample that is like the population so that we can use what we learn about the sample to generalize to the population. What if we already know something about our population, though? How can we make use of that knowledge when constructing our sample? We can replicate known characteristics of a sample by following another probability sampling design, a *proportionate stratified sampling design*. Perhaps we'd like to sample students at a particular college, and we already know students' sex, in-state versus out-of-state residency, and undergraduate versus graduate classification. We can use sex, residency, and classification as our *strata* and select a sample with the same proportions of male versus female, in-state versus out-of-state, and undergraduate versus graduate students as the population. If we determine that 4% of our population are male graduate students from out-of-state and we wanted a sample of 300 students, we'd select (using random sampling or systematic sampling) 12 (300*4%) male graduate students from out-of-state to be in our sample. We'd carry on similarly sampling students with other combinations of these characteristics until we had a sample proportionally representative of the population in terms of sex, residency, and classification. We probably would have gotten similar results if we had used a simple random sampling strategy, but now we've ensured proportionality with regard to these characteristics.

Sometimes, though, proportionality is exactly what we don't want. What if we were interested in comparing the experiences of students who had been homeschooled to students who were not homeschooled? If we followed a simple random sampling design or a proportionate stratified sampling design, we would probably end up with very few former homeschoolers—not enough to provide a basis of comparison to the never homeschooled. We may even want half of our sample to be former homeschoolers, which would require *oversampling* from this group to have their representation in the sample disproportionately high compared to the population, achieved by following a *disproportionate stratified sampling design*. Importantly, this is still a probability sampling design. With some careful math, we can still calculate the probability of any one case in the population being selected for the sample; it's just that for former homeschoolers, that probability would be higher than for the never homeschooled. Knowing these probabilities still permits us to use statistics to estimate parameters for the entire population of students, we just have to remember to make the responses of former homeschoolers count less and the responses of the never homeschooled count more when calculating our parameter estimates. This is done using *weights*, which are based on those probabilities, in our statistical calculations.

One final probability sampling design, *cluster sampling design*, is commonly used to sample cases that are dispersed throughout a broad geographic region. Imagine the daunting task of needing to sample 2,000 parents of kindergarteners from across the United States. There is no master list of kindergarten students or their parents to serve as a sampling frame. Constructing a sampling frame by going school to school across the country would likely consume more resources than the rest of the study itself—the thought of constructing such a sampling frame is ridiculous, really. We could, though, first randomly select, say, 20 states, and then 10 counties within each of those 20 states, and then 1 school from each of those counties, and then 10 kindergartners from each of those schools. At each step, we know the probability of each state, county, school, and kid being selected for the sample, and we can use those probabilities to calculate weights, which means we can still use statistics to estimate parameters. We'll have to modify our definition for probability sampling designs just a bit, though. We *could* calculate the probability of any one case in the population being included in the study, but we don't. Being able to calculate the probabilities of selection for each *sampling unit* (states, counties, schools, kids), though, does the same job, so we still count cluster sampling designs as one of the probability sampling designs. To modify our definition of probability sampling designs, we might say that every case in the population has a known *or knowable*, greater-than-zero probability of being selected for the sample.

Using a probability sampling design is necessary, but not sufficient, if we want to use statistics to estimate parameters. We still need an adequate sample size. How do we calculate an adequate sample size? Do we, say, select 10% of the population? It would be handy to have such an easy rule of thumb, but as it turns out, the size of the population is only one factor we have to consider when determining the required sample size. (By the way, this is probably the most amazing thing you'll learn in this text.) In addition to population size, we also have to consider required level of confidence (something you decide yourself), required level of accuracy (something else you decide), and the amount of variance in the parameter (something you don't get to decide; it is what it is).

As you'd probably guess, the larger the population size, the larger the required sample size. However, the relationship between population size and required sample size is not linear (thus no rule of thumb about selecting 10% or any other percent of the population for your sample). If we have a somewhat small population, we'll need a large proportion of it in our sample. If we have a very large population, we'll need a relatively small proportion of it in our sample. In fact, once the population size goes above around 20,000, the sample size requirement hardly increases at all (thanks again to the central limit theorem and the law of large numbers).

We also have to consider how much the parameter varies. Imagine that I'm teaching a class of 40 students, and I know that everyone in the class is the same age, I just don't know what that age is. How big would my sample size need to be for me to get a very good (even perfect) statistic, the mean age of my students? Think. One! That's right, just one. My parameter, the mean age of the class, has zero variation (my students are all the same age), so I need a very small sample to calculate a very good statistic. What if, though, my students' ages were all over the place—from one of those 14-year-old child geniuses to a 90-year-old great grandmother who decided to finish her degree? I'd be very reluctant to use the mean age of a sample of 3, 4, or even 10 students to estimate the whole class's mean age. Because the population parameter varies a lot, I'd need a large sample. The rule, then: The more the population parameter varies, the more cases I need in my sample.

The astute reader should, at this point, be thinking "Wait a sec. I'm selecting a sample so I can calculate a statistic so I can estimate a parameter. How am I supposed to know how much something I don't know varies?" Good question. Usually, we don't, so we just assume the worst, that is, we assume maximum variation, which places the highest demand on sample size. When we specify the amount of variation (like when using the sample size calculators I'll say more about below), we use the percentage of one value for a parameter that takes on only two values, like responses to yes/no questions. If we wanted to play it safe and assume maximum variation in a parameter, then, we'd specify 50%; if 50% of people in a population would answer "yes" to a yes/no question, the parameter would exhibit maximum variation—it can't vary any more than a 50/50 split. Specifying 0% or 100% would be specifying no variation, and, as it may have occurred to you already, specifying 25% would be the same as specifying 75%.

Very astute readers might have another question: "You've been referring to a required sample size, but required for what? What does it mean to have a required sample size? Isn't that what we're trying to figure out?" Another good question. Given the size of the population (something you don't control) and the amount of variance in the parameter (something else you don't control), a sample size is required to be at least a certain size if we want to achieve a desired level of confidence and a desired level of accuracy, the factors you *do* control. We saw examples of accuracy and confidence previously. We might say "I am 95% percent certain [so I have a 95% confidence level] that the average age of my class is in the 19 to 21 range [so I have a ± 1 year level of accuracy]." A clumsier way to say the same thing would be "If I were to repeat this study over and over again, selecting my sample anew each time, 95% of my samples would have average ages in the range of 19 to 21." Confidence and accuracy go together; it doesn't make sense to specify one without specifying the other. As I've emphasized, you get to decide on your levels of confidence and accuracy, but there are some conventions in social research. The confidence level is most often set at 95%, though sometimes you'll see 90% or 99%. The level of accuracy, which is usually indicated as the range of percentage point estimates, is often set at ±1%, 3%, or 5%. If you're doing applied research, you might want to relax these standards a bit. You might decide that a survey giving you ±6% at an 85% confidence level is all you can afford, but it will help you make decisions better than no survey at all.

So far, I've just said we need to "consider" these four factors—population size, parameter variation, degree of accuracy, and degree of confidence, but, really, we have to do more than just consider them, we have to plug them into a formula to calculate the required sample size. The formula isn't all that complicated, but most people take the easy route and use a sample size calculator instead, and so will we. Several good sample size calculators will pop up with a quick internet search. You enter the information and get your required sample size in moments. Playing around with these calculators is a bit mind boggling. Try it out. What would be a reasonable sample size for surveying all United States citizens? What about for all citizens of Rhode Island? What's surprising about these sample sizes? Play around with different levels of confidence, accuracy, and parameter variation. How much do small changes affect your required sample sizes?

And note the interplay of confidence and accuracy. For any given sample size, you can have different combinations of confidence and accuracy, which will have an inverse relationship—as one goes up, the other goes down. With the same sample, I could choose either to be very confident about an imprecise estimate or to be not-so-confident about a precise estimate. I can look over a class of undergraduates and predict with near certainty that their average age is between 17 and 23, or I can predict with 75% confidence that their average age is between 19 and 20.

It's important to realize what we're getting from the sample size calculator. This is the minimum sample size if we're intending to use statistics to estimate single parameters, one by one—that is, we're calculating univariate statistics. If, however, we're planning to compare any groups within our sample or conduct any bivariate or multivariate statistical analysis with your data, our sample size requirements will increase accordingly (and necessitate consulting statistics manuals).

Calculating a minimum sample size based on the desired accuracy and confidence only makes sense if we're following a probability sampling design. Sometimes, though, our goal isn't to generalize what we learn from a sample to a population; sometimes, we have other purposes for our samples and use *nonprobability sampling designs.* Maybe we're doing a trial run of our study. We just want to try out our questionnaire and get a feel for how people will respond to it, so we use a *convenience sampling design*, which is what it sounds like—sampling whatever cases are convenient. You give your questionnaire to your roommate, your mom, and whoever's waiting in line with you at the coffee shop. Usually, convenience sampling is used for *field testing* data collection instruments, but it can also be used for *exploratory research*—research intended to help orient us to a research problem, to help us figure out what concepts are important to measure, or to help us figure out where to start when we don't have a lot of previous research to build on. We know that we have to be very cautious in drawing conclusions from exploratory research based on convenience samples, but it can provide a very good starting point for more generalizable research in the future.

In other cases, it would be silly to use a probability sampling design to select your case. What if you wanted to observe people's behavior at Green Party rallies? Would you construct a sampling frame listing all the upcoming political rallies and randomly select a few, hoping to get a Green Party rally in your sample? Of course not. Sometimes we choose our sample because we want to study particular cases. We may not even describe our case selection as sampling, but when we do, this is *purposive sampling*. We can also use purposive sampling if we wish to describe typical cases, atypical cases, or cases that provide insightful contrasts. If I were studying factors associated with nonprofit organizational effectiveness, I might select organizations that seem similar but demonstrate a wide range of effectiveness to look for previously unidentified differences that might explain the variation. Purposive sampling is prominent in studies built around in-depth qualitative data, including case studies, which we'll look at in a bit.

When purposively selecting cases of interest, we should take care not to draw unwarranted conclusions from cases *selected on the dependent variable*, the taboo sampling strategy. Imagine we want to know whether local governments' spending on social media advertising encourages local tourism. Our independent variable is social media advertisement spending, and our dependent variable is the amount of tourism. If we were to adopt this taboo sampling strategy, we would identify localities that have experienced large increases in tourism. We may then, upon further investigation, learn they had all previously increased spending on social media advertising and conclude that more advertising spending leads to more tourism. Can we legitimately draw that conclusion, though? It may be that many other localities had also increased their social media advertising spending but did not see an increase in tourism; the level of spending may not affect tourism at all. It's even possible that other localities spent more on social media advertising—we do not know because we fell into the trap of selecting cases on the dependent variable.

We may wish to do probability sampling but lack the resources, potentially making a *quota sampling design* a good option. This is somewhat of a cross between convenience sampling design and the stratified sampling designs. Before, when we wanted to include 12 male out-of- state graduate students in our sample, we constructed a sampling frame and randomly selected them. We could, however, select the first 12 male out-of-state graduate students we stumble upon, survey them to meet our quota for that category of student, and then seek out students in our remaining categories. (This is what those iPad-carrying marketing researchers at the mall and in theme parks are doing—and why they'll ignore you one day and chase you down the next.) We'd still be very tentative about generalizing from this sample to the population, but we'd feel more confident than if our sample had been selected completely as a matter of convenience.

One final nonprobability sampling design is useful when cases are difficult to identify beforehand, like meth users, sex workers, or the behind-the-scenes movers-and-shakers in a city's independent music scene. What's a researcher wanting to interview such folks to do? Post signs and ask for volunteers? Probably not. She may be able to get that first interview, though, and, once that respondent trusts her, likes her, and becomes invested in her research, she might get referred to a couple more people in this population, which could lead to a few more, and so on. This is called (regrettably, I think, because I'd hate to have the term *snowball* in my serious research report) a *snowball sampling design* or (more acceptably but less popularly) a *network sampling design*, and it has been employed in a lot of fascinating research about populations we'd otherwise never know much about.

FORMAL RESEARCH DESIGNS

Simply collecting data is insufficient to answer research questions. We must have a plan, a *research design*, to enable us to draw conclusions from our observations. Different methodologists divvy up the panoply of research designs different ways; we'll use five categories: cross-sectional, longitudinal, experimental, quasi-experimental, and case study.

Cross-sectional research design is the simplest. Researchers following this design are making observations at a single point in time; they're taking a "snapshot" of whatever they're observing. Now, we can't take this too literally. A cross-sectional survey may take place over the course of several weeks. The researcher won't, however, care to distinguish between responses collected on day 1 versus day 2 versus day 28. It's all treated as having been collected in one wave of data collection. Cross-sectional research design is well suited to descriptive research, and it's commonly used to make *cross-case comparisons*, like comparing the responses of men to the responses of women or the responses of Republicans to the responses of Democrats. If we're interested in establishing causality with this research design, when we have to be sure that cause comes before effect, though, we have to be more careful. Sometimes it's not a problem. If you're interested in determining whether respondents' region of birth influences their parenting styles, you can be sure that the respondents were born wherever they were born before they developed any parenting style, so it's OK that you're asking them questions about all that at once. However, if you're interested in determining whether interest in politics influences college students' choice of major, a cross-sectional design might leave you with a chicken-and-egg problem: Which came first? A respondent's enthusiasm for following politics or taking her first political science course? Exploring causal research questions using cross-sectional design isn't verboten, then, but we do have to be cautious.

Longitudinal research design involves data collection over time, permitting us to measure change over time. If a different set of cases is observed every time, it's a *time series* research design. If the same cases are followed over time, with changes tracked at the case level, it's a *panel* design.

Experimental research design is considered by most to be the gold standard for establishing causality. (This is actually a somewhat controversial statement. We'll ignore the controversy here except to say that most who would take exception to this claim are really critical of the misapplication of this design, not the design itself. If you want to delve into the controversy, do an internet search for federally required randomized controlled trial program evaluation designs.) Let's imagine an experimental-design study of whether listening to conservative talk radio affects college students' intention to vote in an upcoming election. I could recruit a bunch of students (with whichever sampling plan I choose) and then have them all sit in a classroom listening to MP3 players through earbuds. I would have randomly given half of them MP3 players with four hours of conservative talk radio excerpts and given the other half MP3 players with four hours of muzak. Before they start listening, I'll have them respond to a questionnaire item about their likelihood of voting in the upcoming election. After the four hours of listening, I'll ask them about their likelihood of voting again. I'll compare those results, and if the talk radio group is now saying they're more likely to vote while the muzak group's intentions stayed the same, I'll be very confident in attributing that difference to the talk radio.

My talk radio experiment demonstrates the three essential features of experimental design: random assignment to experimental and control groups, control of the experimental setting, and manipulation of the independent variable. *Control* refers to the features of the research design that rule out competing explanations for the effects we observe. The most important way we achieve control is by the use of a *control group*. The students were *randomly assigned* to a control group and an *experimental group*. The experimental group gets the "treatment"—in this case, the talk radio, and the control group gets the status quo—in this case, listening to muzak. Everything else about the experimental conditions, like the time of day and the room they were sitting in, were controlled as well, meaning that the only difference in the conditions surrounding the experimental and control groups was what they listened to. This *experimental control* let me attribute the effects I observed—increases in the experimental group's intention to vote—to the cause I introduced—the talk radio.

The third essential feature of experimental design, manipulation of the independent variable, simply means the researcher determines which cases get which values of the independent variable. This is simple with MP3 players, but, as we'll see, it can be impossible with the kinds of phenomena many social researchers are interested in.

Experimental methods are such strong designs for exploring questions of cause and effect because they enable researchers to achieve the three criteria for making causal claims—the standards we use to assess the validity of causal claims: time order, association, and nonspuriousness. Time order is the easy one (unless you're aboard the starship Enterprise). We can usually establish that cause preceded effect without a problem. Association is also fairly easy. If we're working with quantitative data (as is usually the case in experimental research designs), we have a whole arsenal of statistical tools for demonstrating whether and in what way two variables are related to each other. If we're working with qualitative data, good qualitative data analysis techniques can convincingly establish association, too.

Meeting the third criterion for making causal claims, nonspuriousness, is trickier. A spurious relationship is a phony relationship. It looks like a cause-and-effect relationship, but it isn't. *Nonspuriousness*, then, requires that we establish that a cause-and-effect relationship is the real thing—that the effect is, indeed, due to the cause and not something else. Imagine conducting a survey of freshmen college students. Based on our survey, we claim that being from farther away hometowns makes students more likely to prefer early morning classes. Do we meet the first criterion? Yes, the freshmen were from close by or far away before they ever registered for classes. Do we meet the second criterion? Well, it's a hypothetical survey, so we'll say yes, in spades: Distance from home to campus and average class start time are strongly and inversely correlated.

What about nonspuriousness, though? To establish nonspuriousness, we need to think of any competing explanations for this alleged cause-and-effect relationship and rule them out. After running your ideas past the admissions office folks, you learn that incoming students from close by usually attend earlier orientation sessions, those from far away usually attend later orientation sessions, and—uh-oh—they register for classes during orientation. We now have a potential competing explanation: Maybe freshmen who registered for classes later are more likely to end up in early morning classes because classes that start later are already full. The students' registration date, then, becomes a potentially important *control variable*. It's potentially important because it's quite plausibly related to both the independent variable (distance from home to campus) and the dependent variable (average class start time). If the control variable, in fact, *is* related to both the independent variable and dependent variable, then that alone could explain why the independent and dependent variables *appear* to be related to each other when they're actually not. When we do the additional analysis of our data, we confirm that freshmen from further away did, indeed, tend to register later than freshmen from close by, that students who register later tend to end up in classes with earlier start times, and, when we control for registration date, there's not an actual relationship between distance from home and average class start time. Our initial causal claim does not achieve the standard of nonspuriousness.

The beauty of experimental design—and this is the crux of why it's the gold standard for causal research—is in its ability to establish nonspuriousness. When conducting an experiment, we don't even have to think of potential control variables that might serve as competing explanations for the causal relationship we're studying. By randomly assigning (enough) cases to experimental and control groups and then maintaining control of the experimental setting, we can assume that the two groups and their experience in the course of the study are alike in every important way except one—the value of the independent variable. Random assignment takes care of potential competing explanations we can think of *and* competing explanations that never even occur to us. In a tightly controlled experiment, any difference observed in the dependent variable at the conclusion of the experiment can confidently be attributed to the independent variable alone.

"Tightly controlled experiments," as it turns out, really aren't that common in social research, though. Too much of what we study is important only when it's out in the real world, and if you try to stuff it into the confines of a tightly controlled experiment, we're unsure if what we learn applies to the real thing. Still, experimental design is something we can aspire to, and the closer we can get to this ideal, the more confident we can be in our causal research. Whenever we have a research design that mimics experimental design but is missing any of its key features— random assignment to experimental and control groups, control of the experimental setting, and manipulation of the independent variable—we have a *quasi-experimental design*.

Often, randomly assigning cases to experimental and control groups is prohibitively difficult or downright impossible. We can't assign school children to public schools and private schools, we can't assign future criminals to zero tolerance states and more lax states, and we can't assign pregnant women to smoking and nonsmoking households. We often don't have the power to manipulate the independent variable, like deciding which states will have motor-voter laws and which won't, to test its effects on voting behaviors. Very rarely do we have the ability to control the experimental setting; even if we could randomly assign children to two different kindergarten classrooms to compare curricula, how can other factors—the teachers' personalities, for instance—truly be the same?

Quasi-experimental designs adapt to such research realities by getting as close to true experimental design as possible. There are dozens of variations on quasi-experimental design with curious names like *regression discontinuity* and *switching replications with nonequivalent groups*, but they can all be understood as creative responses to the challenge of approximating experimental design. When we divide our cases into two groups by some means other than random assignment, we don't get to use the term control group anymore, but *comparison group* instead. The closer our comparison group is to what a control group would have been, the stronger our quasi-experimental design. To construct a comparison group, we usually try to select a group of cases similar to the cases in our experimental group. So, we might compare one kindergarten classroom enjoying some pedagogical innovation to an adjacent kindergarten classroom with the same old curriculum or Alabama drivers after a new DUI law to Mississippi drivers not bound by it.

If we're comparing these two groups of drivers, we're also conducting a *natural experiment*. In a natural experiment, the researcher isn't able to manipulate values of the independent variable; we can't decide who drives in Mississippi or Alabama, and we can't decide whether or not a state would adopt a new DUI law. Instead, we take advantage of "natural" variation in the independent variable. Alabama did adopt a new DUI law, and Mississippi did not, and people were driving around in Alabama and Mississippi before and after the new law. We have the opportunity for before-and-after comparisons between two groups, it's just that we didn't introduce the variation in the independent variable ourselves; it was already out there.

Social researchers also conduct *field experiments*. In a field experiment, the researcher randomly assigns cases to experimental and comparison groups, but the experiment is carried out in a real-life setting, so experimental control is very weak. I once conducted a field experiment to evaluate the effectiveness of an afterschool program in keeping kids off drugs and such. Kids volunteered for the program (with their parents' permission). There were too many volunteers to participate all at once, so I randomly assigned half of them to participate during fall semester and half to participate during spring semester. The fall semester kids served as my experimental group and, during the fall semester, the rest of the kids served as my comparison group. At the beginning of the fall semester, I had all of them complete a questionnaire about their attitudes toward drug use, etc., then the experimental group participated in the program while the control group did whatever they normally did, and then at the end of the semester, all the kids completed a similar questionnaire again. Sure enough, the experimental group kids' attitudes changed for the better, while the comparison group kids' attitudes stayed about the same (or even changed a bit for the worse). All throughout the program, the experimental group and comparison group kids went about their lives—I certainly couldn't maintain experimental control to ensure that the only difference between the two groups was the program.

Very strong research designs can be developed by combining one of the longitudinal designs (time series or panel) with either experimental or quasi-experimental design. With such a design, we observe values of the dependent variable for both the experimental and control (or comparison) groups at multiple points in time, then we change (or observe the change of) the independent variable for the experimental group, and then we observe values of the dependent variable for both groups at multiple points in time again.

That's a bit confusing, but an example will clarify: Imagine inner-city pharmacies agree to begin stocking fresh fruits and vegetables, which people living nearby otherwise don't have easy access to. We might want to know whether this will affect area residents' eating habits. There are lots of ways we could go about this study, but probably the strongest design would be an *interrupted time series quasi-experimental design*. Here's how it might work: Before the pharmacies begin stocking fresh produce, we could conduct door-to-door surveys of people in two inner-city neighborhoods—one without a pharmacy and one with a pharmacy. We could survey households once a month for four months before the produce is stocked, asking folks about how much fresh produce they eat at home.

(A quick aside: We'd probably want to talk to different people each time since, otherwise, just the fact that we keep asking them about their eating habits, they might change what they eat—an example of a *measurement artifact*, which we try to avoid. We want to measure changes in our dependent variable, *eating habits*, that are due to change in the independent variable, *availability of produce at pharmacies*, not due to respondents' participation in the study itself.)

After the pharmacies begin stocking fresh produce, we would then conduct our door-to-door surveys in both neighborhoods again, perhaps repeating them once a month for another four months. Once we're done, we'd have a very rich dataset for estimating the effect of available produce on eating habits. We could compare the two neighborhoods before the produce was available to establish just how similar their eating habits were before, and then we could compare the two neighborhoods afterward. We might see little difference one month after the produce became available as people became aware of it, then maybe a big difference in the second month in response to the novelty of having produce easily available, and then maybe a more moderate, steady difference in the third and fourth months as some people returned to their old eating habits and others continued to purchase the produce. With this design, we can provide very persuasive evidence that the experimental and comparison groups were initially about the same in terms of the dependent variable, which increases our confidence that any changes we see later are indeed due to the change in the independent variable. We can also capture change over time, which is frequently very important when we're measuring behavioral changes, which tend to diminish over time.

Case study research design is the oddball of the formal research designs. Many researchers who feel comfortable with all the other designs would feel ill equipped to undertake a case study. A *case study* is the systematic study of a complex case that is in-depth and holistic. Unlike the other designs, we're just studying a single case, which is usually something like an event, such as a presidential election, or a program, such as the operation of a needle exchange program. With the other designs, we usually rely on a single data collection method, but with case study research design, we use multiple data collection methods, with a heavy emphasis on collecting qualitative data. In the course of a single case study, we might conduct interviews, conduct focus groups, administer questionnaires, survey administrative records, and conduct extensive direct observations. We make enough observations in as many different ways as necessary to enable us to write a rich, detailed description of our case. This written report is, itself, called a *case study*.

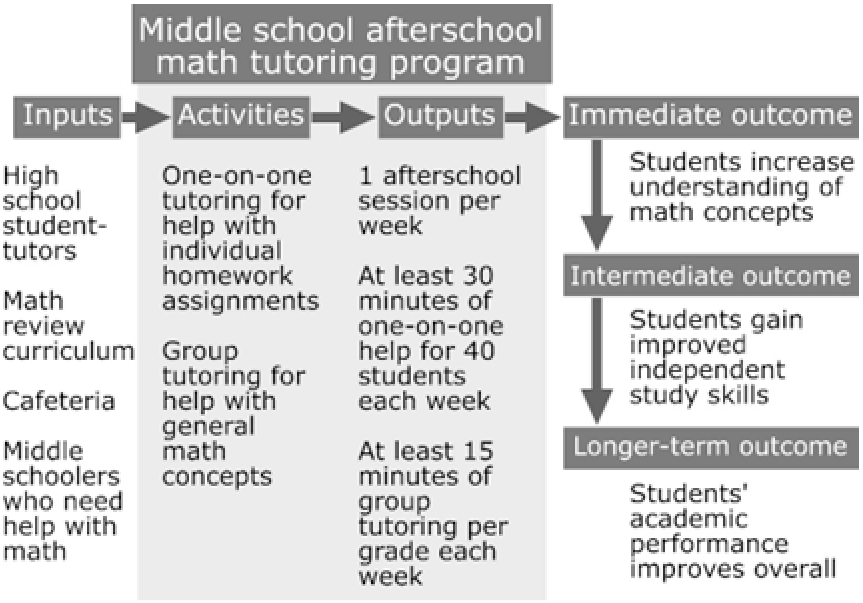
The richness of case studies highlights another key difference between this and the other research designs. The contrast with experimental design is sharpest: If you think about experimental design, its beauty lies in *ignoring* complexity. If I were to randomly assign a bunch of teenagers to experimental and control groups, my express intention would be to ignore all their pimply, hormonal, awkward, exuberant complexity and the group dynamics that would undoubtedly emerge in the two groups. I count on random assignment and experimental control to make all differences between the two groups a complete wash except the difference in the independent variable. With case studies, though, we embrace this complexity. The whole point is to describe this rich complexity, bringing only enough organization to it to make it understandable to people who can't observe it directly—those people who will ultimately read our written case studies.

There are many elaborations on these formal research designs. A few more, along with a system of notation for depicting research designs, are presented in Appendix B.

APPENDIX A: APPLIED RESEARCH AND PROGRAM LOGIC MODELS

You may hear social research referred to as *pure* or *applied*. *Pure research* aims to build knowledge for its own sake; *applied research* aims to be useful for doing things like solving problems, making the most of resources, identifying opportunities for improvement, and planning how to reach a goal. These can be useful distinctions, but they're not mutually exclusive categories. Much pure research is eventually very useful, and much enlightening knowledge is generated in the course of conducting applied research.

When conducting applied research about a program, organization, or policy, models often play the role of theory in the research process. I'll focus here on how models can help generate empirical research questions. The following logic model, for example, depicts how a simple afterschool tutoring program is intended to work.


Middle school afterschool math tutoring program
Inputs
Activities
Outputs
Immediate outcome
High school student-tutors
Math review curriculum
Cafeteria
Middle schoolers who need help with math
One-on-one tutoring for help with individual homework assignments
Group tutoring for help with general math concepts
1 afterschool session per week
At least 30 minutes of one-on-one help for 40 students each week
At least 15 minutes of group tutoring per grade each week
Students increase understanding of math concepts
Intermediate outcome
Students gain improved independent study skills
Longer-term outcome
Students' academic performance improves overall

The inputs include all of the resources for the program (high school student-tutors, curriculum, and the cafeteria) and the demand for the program (middle schoolers who need help with math). The activities are the main actions undertaken by the program, and the outputs are the observable, countable units of service produced. The outcomes depict the chain of intended program results—the ways the program is intended to make the world a better place.

All components of the logic model can generate applied research questions to guide inquiry that could be helpful for the entire program planning and evaluation process. Here are some examples…

Questions to understand and establish the need for the program:

1. How many middle schoolers need help?
2. What are the middle schoolers' academic strengths?
3. What math concepts are especially challenging for the middle schoolers?
4. What are the middle schoolers' study habits?
5. How do the middle schoolers feel about learning math?

Questions about program resources:

6. What tutoring skills do the high school students have?
7. What math knowledge do the high school students have?
8. How much time do the high school students have to commit to the program?
9. Is the cafeteria environment conducive to learning?

Questions about activities and outputs:

10. Are the high schoolers using good tutoring practices?
11. Are the high schoolers following the group tutoring curriculum?
12. Are the middle schoolers staying actively engaged in the tutoring?
13. Are there any barriers to middle schoolers' participation?
14. What do the middle schoolers believe is the most helpful about the program?
15. What do high schoolers think is going well? What concerns do they have?

Questions about outcomes and possible unintended consequences:

16. Are the middle schoolers gaining a better understanding of the targeted math concepts?
17. Are the middle schoolers' grades in math improving?
18. Are the middle schoolers developing better independent study skills?
19. How are the middle schoolers' study habits changing?
20. How is the program affecting middle schoolers' overall academic performance?
21. How is the program affecting middle schoolers' attitudes toward school and learning?
22. How is the program affecting middle schoolers' participation in co-curricular activities?
23. How is the program affecting the high schoolers' educational aspirations?
24. How is the program affecting high schoolers' academic performance?
25. What changes in the students have their parents observed?
26. What changes in the students have their teachers observed?
27. How will the program affect middle schoolers' academic performance next year?

Questions linking activities, outputs, and outcomes:

28. How much time in one-on-one tutoring is sufficient for improving middle schoolers' understanding of the targeted math concepts?
29. Do middle schoolers who participate more often achieve larger gains in academic performance?
30. Which middle schoolers benefit the most from the group tutoring sessions?
31. How does participating middle schoolers' academic performance differ from non- participating students' academic performance?
32. How do the students feel they've changed due to participating in the program?

Involving stakeholders in the design of applied research projects is the most important strategy for producing useful findings. Start by identifying the stakeholders: Who could benefit or suffer based on what we learn? Who can make authoritative decisions based on what we learn? Who will need to approve those decisions? Who will be in charge of implementing changes? These stakeholders can be involved in every stage of the research process. Collaborating with stakeholders, like program managers, on the preliminary step of developing a program model is almost always beneficial for identifying gaps in knowledge or surfacing disagreements over how the program is assumed to function by different stakeholders. These are prime opportunities for developing research questions. Other research questions can be identified by asking stakeholders what decisions they hope to make based on what is learned from the research project. Applied researchers should be certain they have a shared understanding of the meaning of key concepts. (I once spent hours making sure I understood what a program evaluation client meant by *life vision*.) Stakeholders should agree that the operationalizations of those concepts are valid. An entire applied research project will fail if, in the end, a key stakeholder looks at some undesirable findings and dismisses them with "Well, that questionnaire really wasn't a good indicator of our program's outcomes." Stakeholders can provide valuable insight even into data collection plans—they know when kids will be unable to focus because of the school band practicing next door, too drowsy to talk after lunch, and too distracted by the countdown to spring break to bother with your questionnaire. When data have been collected and it's time for data analysis and reporting, stakeholders can participate deliberatively, and then they will be much more inclined to take the findings seriously and use them for policy and program improvement. And if you're conducting applied research, that's the whole point.

APPENDIX B: MORE RESEARCH DESIGNS

This appendix recaps some of the formal research designs covered in the main text and introduces some elaborations on these designs. We'll learn about these designs as applied to program evaluation. Program evaluation is the use of research methods to learn about programs—such as job training programs, dropout prevention programs, substance abuse treatment programs, and so on—with the goals of learning about their effectiveness or how to improve them. I find that students tend to get the idea of using research methods this way very intuitively, so it's a helpful lens for learning about research methods generally. You've all casually evaluated programs a lot—think about why you chose one college over others, why you chose your major, and how you've come up with ideas for how to make your major even better. Program evaluation accomplishes this same kind of thinking, but based on systematic observations using the tools of empirical social science research.

Along the way, we'll also learn the standard notation system for research designs. This system of notation makes it much easier for us to communicate about research designs, so be sure you master this system of notation in addition to learning about the evaluations design themselves.

Our notation will use three letters: R, X, and O. R stands for *random assignment* (and will only be used to depict research designs that use random assignment). X represents our program "happening"—the "intervention" in the terminology of clinical psychology. O stands for *observation*. This refers to observing our outcome indicators. In research methods jargon, X represents the value of the independent variable (IV) that we want to know the effect of, and O represents the act of measuring the dependent variable (DV). So, if we were evaluating a job placement program, X would represent clients participating in the program, and O would represent measuring the key outcomes of that program—whether or not the clients are employed, or maybe their earnings. Program implementation functions as an independent variable (it "happens" to particular people or not), and our outcomes (employment status, wages) function as our dependent variables. The program manager's hope is that the program (IV) will have a positive effect on the outcomes (DV).

We can use these three letters to depict all sorts of research designs. We could start with simple outcome measurement. With this type of evaluation, we make observations (O) of our outcomes just once—at the conclusion of an instance of program implementation (like at the conclusion of a client participating in the program). This should remind you of a research methods design: cross-sectional research design—our observations are made at one point in time with no effort to track change in our DV over time.

We can depict this design like this:

X	O

We read that from left to right: The program happens (X), and then we make our observations (O). Another term for this is *single-group posttest-only* evaluation design. That means we're making observations of just one group (usually people participating in our program, but it could also be, say, stretches of highway in an anti-litter program), and we're measuring out outcomes only after the program.

(That term, *posttest*, like *pretest*, which we will see in a minute, makes it sound like the only way we measure outcomes is by administering tests—fortunately, that's very much not the case, but it is an unfortunate implication of the term. You can use other terms, like *before and after* to get around that bit of confusion, but we'll go with these terms for now.)

This is a very simple evaluation design, and it's very common. Sometimes, it's sufficient because we can confidently attribute the outcomes we observe to the program. Imagine a program in which employees attend a one-hour workshop on how to use the new campus intranet. There's no way they would have had that knowledge beforehand, so if we observe indicators of their knowledge of the system after the program (like on a quiz—always makes for a fun way to end a workshop!), we can be quite confident that they gained that knowledge during the workshop.

Often, however, the single-group post-only design is weak because we can't know that the observed outcomes are truly due to the program. (This would be weak internal validity, remember, in research methods jargon.) Imagine, instead, a 3-month program of weekly, one- hour workshops intended to improve employees' workplace communication skills. You could use the simple X O design, but what if you observed indicators of excellent workplace communication skills? How confidently can you attribute those outcomes to the program? How do you know the participants didn't already have strong communication skills? Or that they started with good communication skills, and now they have just slightly better communication skills? Or that they started with excellent communication skills, and now their skills are actually worse because they're so afraid of messing up? The X O design can't let us explore any of those possibilities.

There are two main approaches (and many, many elaborations on these two approaches) to strengthening the internal validity of our evaluations: (1) making observations over time, and (2) making comparisons. Let's start with making observations over time. That should call to mind our longitudinal designs—time series and panel. We'll usually be using panel designs.

For example, our workplace communication workshop participants might take a pretest—a measure of our outcome before the program and then a posttest—again, a measure of our outcome—after the program. That way, we can track changes in the individual participants' levels of communication skills over time. This is a *single-group pretest/posttest design*, depicted like this:

O_1	X	O_2

Notice that we're now designating our observations with subscript numbers to help us keep them straight.

The single-group pretest/posttest design is a big improvement over the single-group posttest- only design. We can now see if our outcome indicators actually change from before to after the program. This is also a very common evaluation design, and, like the X O design, it may be adequate if you can confidently attribute the changes you observe to the program and not to some other factor. If we did see improvements in our participants' workplace communication skills, we'd probably be pretty confident in attributing those improvements to our program.

Let's imagine still another scenario, though. What if we're evaluating a 12-week youth development program that involves weekly small group meetings with the goal of helping middle schoolers improve their self-image? A single-group pretest/posttest design would be better than nothing, but what if we did see improvement in our self-image indicators? How would we know that the program had made the difference? What if improvements in self-image just tend to happen naturally as kids become more acclimated to their middle schools and make new friends and so on? Or what if something else happened during the program—like what if they all happened to start doing yoga in PE, and that made the difference in their self-image? How do we know that these kids' self-images wouldn't have improved even without the program? To answer those questions, we need to use that second strategy for strengthening the internal validity of our evaluations: making comparisons.

Here's where we come to the evaluation design that, as we've already learned, is considered the gold standard in evaluation design: experimental design. Here's how we depict the classic experimental design:

R	O_1	X	O_2
R	O_3		O_4

Now we have two rows, which indicates that we have two groups. The top row depicts the experimental group, also called the treatment group. In a client-serving program, this would be a group of people participating in our program. The second row depicts the control group. This is a group of people who do not participate in the program—they receive no services or just whatever the status quo is.

The Rs indicate that the clients participating in our evaluation were randomly assigned to the experimental and control groups. Remember, random doesn't mean haphazard. Random assignment means that all of our cases—usually the people participating in our evaluation—had an equal probability of being assigned to the experimental group or the control group. This is really important because it means that, with a large enough number of participants, we can figure that the two groups were, on average, pretty much the same. They're the same in terms of things we might think about—like motivation for change or pre-existing knowledge, and they're also the same even in terms of things we don't ever think about. The only difference, then, between the two groups is that the experimental group participates in the program and the control group does not.

The features of the experimental design give us a lot of confidence in attributing changes in outcomes to the program. We can see before-to-after change by comparing O_1 to O_2, and we can rule out the possibility that the change would have occurred even without the program by observing the control group's outcome indicator changes from O_3 to O_4. This is key—because of random assignment, we can assume that the two groups started out pretty much the same in terms of the outcome we're interested in and even in terms of everything else that might affect outcomes—things like their motivation or pre-existing knowledge. We can even double-check some of this by comparing O_1 to O_3, which we'd expect to be close to the same. And if there would have been some "natural" improvement in the outcome even without the program, we can account for that.

This is accomplished by calculating the *difference in differences*—that's $[(O_2-O_1)-(O_4-O_3)]$—very literally the difference between the two groups of their differences from before to after the program.

Let's look at some numbers to help that make sense. Let's say we're measuring our youth development program's effect on our participants' self-image using some kind of an assessment that gives a score from 0 to 100, and that we observe these average scores for our experimental and control groups before and after the program:

R	60	X	80
R	60		70

Here, I've substituted the two groups' average pretest and posttest scores for the O_1, O_2, O_3, and O_4. First, note that our random assignment worked—our average pre-program outcome measures are the same for our experimental and control group. (In real life, these numbers wouldn't be exactly the same, but they should be close.)

So, did our program work? Well, the program participants' scores increased by an average 20 points, so that's good. But our control group's scores increased by an average 10 points, even without participating in the program. What would be our measure of the program's effectiveness, then? We calculate the difference in differences—we calculate the change for the control group and subtract that from the change for the experimental group: 20 minus 10, or 10 points. We can be very confident, then, that our program accounted for a 10-point improvement in our participants' self-image scores.

We can also see how the experimental design is a big improvement over the other designs. Imagine we had used a single-group posttest-only design:

X	80

We'd be pleased to see a nice, high average outcome score, but we wouldn't be very confident at all in attributing that score to our program. If we used a single-group pretest/posttest design:

60	X	80

... we'd know that our outcome measures had, on average, increased during the program. We'd be very mistaken, though, to attribute this entire increase to our program—something we wouldn't know if we hadn't had the control group for comparison.

There are lots of variations on experimental designs. You might be comparing two different program models instead of comparing a program to no program, which we could depict like this:

R	O_1	X_1	O_2
R	O_3	X_2	O_4

… Now with two experimental groups participating in two different programs, represented by the two Xs, instead of one program and one no-treatment control group.

If you're concerned about testing artifacts—the possibility that the act of taking the pretest might help your participants score better on the posttest, you can explore that possibility with a Solomon 4-group design:

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

Pause for a moment and think about how you would go about looking for a testing artifact. Which observations, or pre-to-post differences would you compare?

OK. Hopefully, you understand why experimental designs are considered the gold standard for evaluating program's effectiveness. They use both strategies for strengthening the internal validity of our designs—we can measure change over time, and we can make good comparisons. Random assignment means that we can be very confident in our comparisons because the only difference between our experimental group and control group is the program, so we can attribute any differences we observe in their outcomes to the program.

Very often, though, experimental designs aren't feasible. A program might be a full coverage program, meaning that everyone who is eligible participates, so there's no viable control group. Or maybe it poses too great an ethical dilemma to withhold services from the control group (though maybe you can overcome that by providing services to the control group after the evaluation). Or maybe it's just too complicated or expensive—very common problems with experimental designs. If these problems cannot be overcome, then a second-best is often a quasi-experimental design.

There are many, many types of quasi-experimental designs. One of the thickest books on my bookshelves is nothing but an encyclopedia of quasi-experimental designs. Obviously, we're not going to cover all of those, but they all have one thing in common: These evaluation designs are all trying to get as close as possible to experimental design while creatively overcoming whatever obstacles keep us from carrying out an experiment in the first place. For the most part, I'm going to leave it at that—all of these quasi-experimental designs are creative solutions to overcoming challenges to carrying out experimental designs. Here's the most common example, though …

If our basic experimental design looks like this:

R	O_1	X	O_2
R	O_3		O_4

Then a very basic quasi-experimental design looks like this:

O_1	X	O_2
O_3		O_4

This is called a *nonequivalent comparison group design*. All we've done is taken away random assignment. Instead of random assignment, we've used some other way to come up with our comparison group (which, recall, we must now call a comparison group, not a control group— the term *control group* is reserved for when we've used random assignment). Maybe we found a similar group—like a class of students in study hall instead to compare to the class of students participating in our program.

However we found our comparison group, the goal is to have a comparison group that is as similar to our experimental group as possible—just like a true control group would have been. This can be very, very tricky.

One big problem is what's called *self-selection bias*, which we considered briefly before. If kids volunteered to participate in our program, meaning they self-selected into our program, then they probably tend to be different somehow than the average non-participant. If we just choose a bunch of other kids to be our comparison group, then, they're probably not really a very good comparison group. We'd need to figure out some way to find a comparison group that had similar motivations—like a group of kids who volunteered for the program but couldn't participate because of scheduling conflicts or had to be placed on a waiting list because we had too many volunteers. There are a lot of other ways of dealing with this problem and other problems you may encounter when designing a quasi-experimental evaluation, but we're going to leave our discussion there, and you can learn more about quasi-experimental designs on an as-needed basis when you're working on your own evaluations.

Sometimes, you're going to be stuck with a single-group design, like in the full coverage scenario I mentioned earlier or when you otherwise just can't develop a strong comparison group. In that case, we do have some strategies for improving the single-group design beyond the basic X O or O_1 X O_2.

I bet you can learn one way just by looking at the notation. See if you can interpret this:

O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8

As I'm sure you can figure out, here we have a panel design with multiple pretests and multiple posttests.

This is called an *interrupted panel design* (or, if we're observing different cases over time, an *interrupted time series design*—recall the difference between panel and time series designs). This way, we can have a sense of any changes that are ongoing before the program and take that into account when interpreting our outcomes measures after the program. If those middle school students' self-images were gradually improving before the program and then continued to gradually improve after the program, we'd be very cautious in attributing the changes to our program—something we may have missed if we'd done a simple before-and-after design.

By the way—to back up a little bit—we can have a really strong quasi-experimental design by combining the interrupted panel design and the nonequivalent comparison group design like this:

O_1	O_2	O_3	X	O_4	O_5	O_6
O_7	O_8	O_9		O_{10}	O_{11}	O_{12}

This is called a *multiple interrupted panel design* or *multiple interrupted time series design*. Pause for a moment to make sure you understand what we're doing here and why it would be such a strong evaluation design.

Back to improving the single-group design. We can also do something that's a bit harder to depict with our notation: Make some outcome measures *during* the program itself. These are, rather inelegantly, called "during" measures, and you'll even see these designs referred to as *single-group before-during-during-during-after designs*. That's pretty awful sounding, but very descriptive, too! I've seen one stab at depicting this design like this:

O_1	[X ...	O_2	O_3	O_4	... X]	O_5

... with the brackets suggesting that the observations are taking place while the program is underway. If we did this with our 12-week youth development program, we could see if there were any changes in response to particular parts of the program. This design gives us the opportunity to associate changes in outcomes with specific events in the program, which gives us a lot more confidence in attributing changes in outcomes to the program than a simple before-and-after design.

One final option is a *dose-response design*. This design might be depicted just like the other single-group designs, but in the previous designs, we've treated the independent variable as a dichotomy—either the program happened or it didn't. With a dose-response design, instead, we treat the independent variable as a continuous variable—as a program that can happen a little or a lot. In our youth development program, for example, some kids may participate in the program for 6 hours, others may participate for 10 hours, others may participate for 12 hours, and so on. We can make the most of this variation in the independent variable to determine if "more" program results in better outcomes. We'd have to make sure we're not accidentally seeing the results of something else—like the kids' motivation to participate—but this design can give us another opportunity to determine if changes in outcomes really can be attributed to the program, even with just a single-group design.

Finally, we can also use a case study approach for our program evaluation design. Case studies, with their multiple sources of data and multiple data collection methods, create a very in-depth, holistic description of the program. This is an especially helpful approach if your evaluation is intended to pursue a formative purpose—the purpose of learning how to improve a program.

APPENDIX C: INFERENTIAL STATISTICS

I'm including this appendix as a very general refresher on inferential statistics for students who may be a bit fuzzy on the concept. This should also help students make connections between what you learn in a statistics course and research methods concepts. Some of our research methods concepts are included in this review as well to help you make those connections. We'll also review the uses and limitations of *p*-values and consider two strategies for overcoming those limitations.

Inferential statistics is the branch of statistics that helps us use characteristics of a sample to estimate characteristics of a population. This contrasts with *descriptive statistics*, which are those statistical tools that describe the data at hand without attempting to generalize to any broader population.

We're very familiar with examples of inferential statistics. For example, news reports commonly report the results of public opinion surveys, like the presidential approval rating. The surveyors—like the Gallup organization or CNN—randomly select maybe 1,500 adults from across the country and ask them whether or not they approve of the president's performance. They might learn that, say, 45% of those surveyed approve of the president's performance. The point, though, is to estimate what percentage of all adults support the president, not just the 1,500 adults they talked to. The use the 45% approval rating as an estimate of all adults' approval rating.

Let's use this example to learn (and review) some vocabulary:

Population: The population is the entire set of cases that we want to learn about. In our example, the population is all of the country's adults. Note, however, that the population doesn't have to be people. We could want to learn about a population of counties, a population of Supreme Court decisions, a population of high schools, or a population of counseling sessions.

Sample: The sample is the set of cases that you actually collect data for. In our example, the sample is the 1,500 adults actually surveyed.

Statistic: Obviously, we've seen this term before (like at the top of this page!), but here, we're using the term *statistic* in a narrower sense of the term. A statistic is a quantified characteristic of a sample. A quantified characteristic could be a mean, median, mode, frequency, standard deviation, or any number of other measures. In our example, 45% is a statistic. It's a characteristic of the sample of 1,500 adults who were surveyed.

Parameter: A parameter is a quantified characteristic of the population. We usually don't know the parameter—that's why we're collecting data from a sample. Our statistics, then, are used to estimate parameters. In our example, we don't know the parameter we're interested in. We don't know the percentage of all adults who approve of the president's performance. We just know the statistic, so we use that to estimate the parameter. We know it's very unlikely that our statistic is exactly equal to the parameter, but it's our best estimate. If we had taken a different sample, we would have gotten a different statistic, even though the parameter was exactly the same.

Sampling frame: The sampling frame is a list. It's the list that we choose our sample from. Ideally, the sampling frame would include every case in the population. In our example, the ideal sampling frame would be a list of every adult in the United States and their phone numbers. Obviously, no such list exists, so the pollsters have to come up with another strategy.

Sampling strategy: The sampling strategy is the set of rules followed in selecting the sample. A very common sampling strategy is simple random sampling. In simple random sampling, every case in the population has an equal (greater than zero) probability of being selected for the sample. In our example, if we could take the name of every adult, write them on index cards, dump all the index cards in a gigantic hat, mix up the cards really well, and then draw out 1,500 cards, we would have used a simple random sampling strategy. Every case in our population (that is, every adult in our country) would have had an equal probability of being selected for our sample. We learned about other sampling strategies earlier.

Level of confidence and *level of accuracy*: Two terms, but we have to talk about them at the same time. When we use a statistic to estimate the corresponding parameter, we have to report how confident we are in that estimate. In our example, we might see a news report like *45% of American adults approve of the president's performance*, and then in the fine print, *95% level of confidence, ±3%*. That fine print means that if we were to repeat this survey again and again and again at the same time but with a different sample each time, we'd expect the statistic to fall between 42% and 48% in 95% of those surveys. Put another way, we're 95% sure that the parameter is somewhere between 42% and 48%. (This is due to the central limit theorem, which tells us that statistics, when calculated from the same population again and again and again, many, many times, will follow a normal distribution. This is amazing stuff. Order out of chaos! It's what makes most inferential statistics work. But back to confidence and accuracy…) In that statement from the news report, 95% is (obviously) the level of confidence, and ±3% is the level of accuracy. Here's why we can only talk about these at the same time: Using our same survey of 1,500 adults, if we want to be more confident, like 99% confident, we'd have to be less accurate in our estimate, like maybe ±10%. (Note that ±10% is *less* accurate than ±3% because it's less precise—don't be fooled by the bigger number.) So, using the same data, we might say we're 99% confident (almost positive!) that the population's presidential approval rating is somewhere between 35% and 55%. Not very impressive, right? It's easy to be really, really certain about a really, really imprecise estimate.

We often use inferential statistics to estimate measures of relationships between variables in the population. For example, we might want to know if men and women have different average presidential approval ratings. We could look at our sample data of 1,500 adults, which might include 750 men and 750 women. We could find in our sample that 43% of the men approve of the president's performance, but 47% of the women approve of the president's performance. Here's the thing: Even if men's and women's presidential approval ratings are exactly the same in the population, we wouldn't expect for them to be exactly the same in our sample—that would be an amazing coincidence. We're interested in knowing whether the difference between men and women in our sample reflects a real difference in the population. To do that, we'll conduct what's called *hypothesis testing*.

We'll imagine that there is no relationship between our two variables—gender and presidential approval—in the population. We're just imagining that—we don't know (that's why we're collecting and analyzing data!). We'll then consider our sample data—men's 43% approval rating and women's 47% approval rating—and ask, *What's the probability that we would see that big of a difference between the men and women in our sample if there's really no difference between men and women in the whole population?* Put another way, we're asking *What's the probability that we're observing this relationship between the two variables (gender and presidential approval) if there's really no relationship in the population?* If that probability is really low, we'll reject the idea that there's no relationship and say we are really confident that there most likely is a relationship between the variables in the population. If that probability isn't low enough to satisfy us, we'll say we don't have evidence to reject that idea, so we'll assume there's no relationship between the variables in the population until we get evidence that there is. Our initial assumption that there is no relationship between the variables in the population is called the *null hypothesis*.

The idea that there is a relationship between the variables in the population is called the *alternative hypothesis*. It's the alternative to the null hypothesis suggested by our sample data that we're interested in testing. It's sometimes called the *research hypothesis*.

Statistics is a very cautious field, so we tend to require a high standard of evidence before we reject the null hypothesis and accept the alternate hypothesis. Most often, we'll reject the null hypothesis and "believe" our sample data if there's no more than a 5% chance that we're rejecting the null hypothesis when we really shouldn't. Put another way, we'll reject the null hypothesis if there's no more than a 5% chance that the results we see in our sample data are just due to chance. Sometimes, people will use a 1% or 10% standard, but it's always a pretty conservative standard so that we're very confident in the conclusions we draw about the population from our sample.

Even with such a high standard for evidence, though, there's still a chance that our conclusions are wrong. That's the risk we take if we want to use sample data to draw conclusions about the whole population. If we reject the null hypothesis when we shouldn't have, we've committed what's called a *Type I error*. If we fail to reject the null hypothesis when we should have, we've committed a *Type II error*. In other words, if we conclude from our sample data that there really is a relationship between our variables in the population when there really isn't, we've committed a Type I error; if we conclude from our sample data that there is no relationship between our variables in the population when there really is, we've committed a Type II error.

If you back up two paragraphs, you may notice that I didn't use the term *p-value*, but if you recently took a statistics course, I'm sure it rings a bell. The precise meaning of this term is almost comically debated among statisticians and methodologists. I'm not willing to enter the fray, so I'm going to totally cop out and quote Wikipedia. (Please don't tell your professor.) Here you go:

> "In statistical hypothesis testing, the *p*-value or probability value is the probability of obtaining test results at least as extreme as the results actually observed, assuming that the null hypothesis is correct. A very small *p*-value

> means that the observed outcome is possible but not very likely under the null hypothesis, even under the best explanation which is possible under that hypothesis. Reporting *p*-values of statistical tests is common practice in academic publications of many quantitative fields. Since the precise meaning of *p*-value is hard to grasp, misuse is widespread and has been a major topic in metascience." (https://en.wikipedia.org/wiki/P-value, retrieved July 10, 2020)

That's a good definition. The debate has to do with how we tend to forget what, exactly, we're comparing the observed outcome (the result of our statistical analysis) to. Honestly, the important thing to remember is that a very small value *p*-value—again, less than 0.05 is a common convention—means the results we get from our statistical analysis probably represent a "real" relationship in the population, not just a fluke of our data analysis. I'm going to leave it at that, but if you want to have some fun, delve into the debates over *p*-value interpretation.

I do, however, want to make the case that *p*-values are important, but insufficient, for drawing conclusions from our statistical analysis. This is emphasized in introductory statistics courses much more often than it used to be, but I'll take the opportunity to make the point here just in case you haven't encountered it before.

Let's start by considering this question: *Why aren't p-values enough?* We use *p*-values as a measure of the statistical significance—a measure of how likely or unlikely it would be to get the results we got (like from correlation or a *t*-test) if, in fact, there were no relationship or difference—whatever we're testing for—at all (and if all the real data in the population look like what we assume they look like, such as being normally distributed). (That convoluted last sentence gives you a sense of what the *p*-value interpretation debates are about!) *P*-values let us draw conclusions like *It's really likely that our finding is a fluke; we'd probably get a totally different result with a different sample* and *Our finding is almost definitely not a fluke; it almost definitely represents a real relationship in the population*. Note two things: (1) These conclusions don't say how strong the relationship is, just that it's flukey or not, and (2) *p*-values are extremely sensitive to sample size. It's easy to get a statistically significant finding for a really weak relationship if we have a big enough sample. *P*-values are important but insufficient.

We need to do additional analysis, then. I'll commend two tools to you: emphasizing confidence intervals and calculating effect sizes.

We've already learned about confidence intervals when we talked about the degree of accuracy in the section about sampling and then again up above. A *point statistic* alone can connote an unwarranted degree of precision. It's more honest to report confidence intervals whenever we can—to say, for example, that we're 95% sure the average weekly hours spent studying in the population of students is between 10 and 20 rather than just reporting the point statistic, a mean of 15 hours.

Effect sizes may be new to you, so we'll spend more time here. *Effect sizes* are what they sound like—a way to gauge "how big" the effect of an independent variable is on a dependent variable. They can also be a way to gauge the strength of non-causal relationships. There are many measures of effect size. There are different effect sizes for the various statistical tests, and each of the various statistical tests usually has several different effect sizes for you to choose from.

We're going to learn about effect sizes in general by learning about one specifically: Cohen's *d*. This is a very widely used measure that gives us an effect size when we're comparing the means of two groups or the means of the same group in before-and-after measures. Sound familiar? If you've already taken a statistics course, this should call to mind *t*-tests, and, yes, Cohen's *d* is often coupled with *t*-tests. The *t*-test gives us the *p*-value, our measure of statistical significance, and Cohen's *d* gives us the rest of the information we want, the effect size.

There are a couple of variations of Cohen's *d*. We're going to use the simplest and most widely used version. It uses standard deviation as a measuring stick; you can interpret Cohen's *d* as the number of standard deviations of difference between two means. (If you haven't taken a statistics course yet, just keep skimming for the general idea and come back here once you've taken that course.) Since you're calculating means from two groups, we're faced with the question of which group's standard deviation to use. We dodge the question by just lumping both groups' data together for calculating the standard deviation, then called the *pooled standard deviation*. The formula for Cohen's *d* is:

[(group 1 mean) – (group 2 mean)]/(pooled standard deviation)

That's just the difference in the two groups' means divided by the standard deviation for both groups lumped together.

Which group should be group 2 and which should be group 1? If we're doing a before-and-after analysis, you'd want to subtract the "before" group's mean from the "after" group's mean so that increases in measures from before to after would yield positive effect sizes (and decreases would yield negative effect sizes—yes, that's a thing). You could think of that effect size formula as:

[(the "after" group's mean) – (the "before" group's mean)]/(pooled standard deviation)

If you're comparing two groups' means on a dependent variable to determine the effect of an independent variable, you need to consider what value of the independent variable you want to know the effect of. If you were evaluating the effect of a program with an experimental design, you would deliver the program to one group of people and not deliver the program to a second group of people. Recall, these groups are called the *experimental group* and *control group*, respectively. Your independent variable could be called *whether or not someone participated in the program* (a little wordy, but clear enough!), and your dependent variable would be your measure of the program's effectiveness. In this situation, you'd want to subtract the control group's DV mean from the experimental group's DV mean so that if the program has a positive effect, the effect size is positive (and if the program has a negative effect, the effect size is negative). You could think of that effect size formula as:

[(the experimental group's mean) – (the control group's mean)]/(pooled standard deviation)

Here's an example: Let's say we want to measure the effectiveness of a math tutoring program. We do this by giving a group of students a math test, then we enroll that same group of students in the math tutoring program for 12 weeks, and then we give that same group of students the math test again. Here's the data we gather:

Mean score on the math test before the tutoring program:	68
Mean score on the math test after the tutoring program:	84
Standard deviation of all the tests (before and after):	19

We'll use the formula we looked at above for before-and-after scenario and plug in those numbers:

(84 – 68) / 19
= 0.84

Our effect size, as measured by Cohen's *d*, then , is 0.84.

Here's another example: Let's say we're going to measure the effectiveness of that math tutoring program, but we're going to do that by randomly assigning one group of students to participate in the program and another group to not participate in the program. (We randomly assign them so that the two groups are as similar to each other as possible, except one is participating in our program, but the other isn't. That way, if there's a difference in the two groups' math test scores, we can confidently attribute that difference to the program instead of something else, like the students' motivation or knowledge.) We enroll the first group (the experimental group) in the tutoring program for 12 weeks. We let the second group (the control group) just go about doing whatever they would have done anyway. At the end of the tutoring program, we give both groups a math test. Here's the data we gather:

Experimental group's mean score on the math test:	80
Control group's mean score on the math test:	72
Standard deviation calculated based on all the math tests:	18

We'll plug those numbers into our formula for Cohen's *d*:

(80 – 72) / 18
= 0.44

Our effect size, as measured by Cohen's *d*, then , is 0.44.

Cohen (the guy who came up with this measure) suggested some rules-of-thumb for interpreting effect sizes:

d = 0.2 is a small effect
d = 0.5 is a medium effect
d = 0.8 is a large effect

Cohen, himself, though, emphasized that these are just rough guidelines and that we would be better off comparing the effect sizes we obtain to what other studies get in similar situations to get an idea of the range of typical scores and what might be considered "small" or "large" in the context of those similar studies. Really, though, most people just kind of blindly apply the rules of thumb.

Notice one other benefit of Cohen's *d*: We could compare evaluations of the same program that use different measures of effectiveness. For example, we could compare findings of a 2001 evaluation of The Math

Tutoring Program that used the Fraser Test of Math Ability as the effectiveness measure to a 2010 evaluation of The Math Tutoring Program that used the Wendell Math Aptitude Test as its measure of effectiveness by comparing their Cohen's *d* statistics. This has become a very common and fruitful application of Cohen's *d* and similar effect size statistics.

APPENDIX D: ELABORATION MODELING

There are different ways to introduce control variables into the analyses of causal relationships. One method is to use *elaboration models* (also called the *elaboration paradigm*, but that sounds a bit big-for-its-britches because it's really a very simple tool). We'll look at this in the context of bivariate (one independent variable, one dependent variable) statistical analysis. (Another way to introduce control variables is to use multiple regression, and there are still other techniques for specific types of bivariate statistical analysis.) The same logic can be applied to qualitative data analysis as well.

An elaboration model is fairly simple. If we introduce a control variable, we want to measure the effect of the independent variable (IV) on the dependent variable (DV) while *controlling for* the control variable (CV). Other phrasings are helpful for understanding what we're after:

What is the effect of the IV on the DV variable *holding the CV constant*?

What is the effect of the IV on the DV variable *independent of the influence of the CV*?

What is the effect of the IV on the DV variable, *regardless of the CV*?

For example, we might observe that men make higher wages than women, and we find this to be a statistically significant relationship using a *t*-test to compare men's and women's average wages. Someone might challenge that finding, saying that there's a third variable at play: Years in the workforce. Women are more likely to take time off for raising children, so maybe they tend to make less money because they haven't put in as much time in the workforce. Does the original finding hold up to this challenge? We'd want to see if men make higher wages than women, controlling for years in the workforce. Our IV is gender, our DV is wages, and our CV is years in workforce. We could test the influence of this CV on our causal relationship of interest by asking: What is the effect of gender on wages, controlling for years in the workforce? Put differently,

What is the effect of gender on wages, *holding workers' years in the workforce constant*?

What is the effect of gender on wages, *independent of the influence of workers' years in the workforce*?

What is the effect of gender on wages, *regardless of workers' years in the workforce*?

An elaboration model applies the "holding the CV constant" phrasing quite literally. To investigate this question, we could divide our workers into, say, three categories, based on their values for the control variable: <6 years in the workforce, 6 – 10 years in the workforce, and >10 years in the workforce. Then, we could measure the relationship between sex and wages *within* each of those three levels. That would be three separate *t*-tests: One *t*-test for just the <6 years group, one for just the 6 – 10 years group, and one for just the >10 years group. We would be measuring the relationship between our IV and DV three times, while literally holding the CV constant each time.

What might we learn from applications of elaboration models?

The control variable may have *no influence* on the causal relationship: If the original wage gap persists throughout the three *t*-tests, we would conclude that the men make higher wages than women, controlling for years in the workforce.

The control variable may *wholly explain away* the purported causal relationship, meaning it was a spurious relationship to begin with: If the wage gap disappears throughout the three *t*– tests, we would conclude that there is no relationship between sex and wages when controlling for years in the workforce and that the simple bivariate relationship between sex and wages is spurious. Sex and wages are both related to years in the workforce, but they are not directly related to each other.

The control variable, quite often, will *partially explain away* the causal relationship under investigation, meaning that some, but not all, of the relationship between the IV and DV is really due to both of them being related to the CV. If the three *t*-tests reveal that men have higher wages than women, but to a lesser degree than in the original *t*-test conducted with the entire sample at once, we would conclude that there is, indeed, a wage gap, but part of the wage gap is attributed to differences in men's and women's years in the workforce.

The control variable may help to better *specify* the relationship between the IV and DV: If the *t*– tests reveal no wage gap among the <6 year workers, a moderate wage gap among the 6 – 10 year workers, and a larger wage gap among the >10 year workers, the control variable has helped us describe the relationship between sex and wages with better specificity.

In crazy, uncommon cases, the control variable may have a *suppressor effect*, revealing a stronger relationship between the IV and CV or even changing the direction (direct to inverse) of the relationship between the IV and CV. If our three *t*-tests revealed that, within each of the groups, women had higher wages than men, we would conclude that we need to spend more time with our data to figure out the complex causal relationships at work between sex, wages, and years in the workforce!

APPENDIX E: PROMOTING EQUITY IN AND WITH SOCIAL SCIENCE RESEARCH

You have come to understand that social research is not a value-free enterprise. Our values shape our choice of research topics, our methodological choices, and the meaning we construct from the results of our data analysis. Our research can also be used to pursue values, such as by conducting applied research to optimize values like effectiveness or efficiency. Equity, or fairness, is a value that deserves careful attention. Our methodological choices and what we learn through research can promote equity or inadvertently perpetuate inequity. The most obvious way research can focus on equity is in the selection of our research questions. Social research is commonly used to explore questions about disparities among different racial and ethnic groups, geographic regions, genders, and socioeconomic groups and to identify ways to improve equity. I just entered "racial disparities" in Google Scholar and found examples of social science research seeking to describe and explain racial disparities in education, health care, and criminal justice just on the first page of results. There are other, perhaps less obvious, ways that our research choices can affect equity, whether our research question is directly about equity or not. Below, I offer some principles and practices to consider as we plan and conduct our own research with the value of equity in mind.

(1) *Pursue research questions with the goal of describing and explaining inequities and identifying possible remedies.* I'm including this first just in case you skipped the paragraph above and jumped straight to the numbered list. (I do that a lot.) Go back and read that paragraph. This is the most important strategy for pursuing the goal of equity with research.

(2) *Disaggregate data to identify inequities*. Even if our research project isn't about inequity *per se*, we can still take the opportunity to look for evidence of equity and inequity. This is very common in program evaluations. The primary goal of an evaluation of an afterschool tutoring program may be to determine if students' academic performance improves due to participation in the program. We may take the opportunity, though, to disaggregate our data to ask comparative questions from an equity perspective: Does the program work equally well for students of different genders? Different races? Different ages? For native and non-native English speakers? Follow-up research questions could explore why we do or do not see disparities, which could help people leading this tutoring program and similar programs to improve or sustain equitable outcomes.

(3) *Conduct within-group analysis*. I think the most overlooked opportunity for conducting research from an equity perspective is to examine variation in outcomes within groups. Imagine that we conduct our evaluation of the afterschool tutoring program, disaggregate our data, and discover that native English speakers see improved academic performance as a result of participating in the program, but non-native speakers do not. A next step could then be to look at variation in outcomes *within* the group of non-native speakers. Most likely, we will learn that, while they benefit less than the native speakers on average, there is still variation in

learning outcomes among the non-native speakers. Some of these students probably benefit from the tutoring program more than others. We may be able to identify factors that help explain that variation. Did the students for whom the program was helpful have tutors who also spoke their native language? Did these students seek help with one subject more often than another? Do these students have different levels of parental support? By exploring within-group variation, we are able to go beyond simply identifying disparities to identifying possible strategies for reducing disparities.

(4) *Be thoughtful about demographic control variables*. This appendix follows the appendix on elaboration modeling in hopes that you already have a good grasp of the role of control variables in our research. (If you are unsure about why we use control variables, reading about elaboration modeling first is a good idea.) Demographic factors are often included in research designs as control variables. This is, in itself, fine and often a good idea. We must, though, take care in how we interpret our findings. Imagine reading this interpretation of multiple regression results in a journal article reporting the outcomes of a job training program evaluation:

For every additional month of job training, the model predicts participants' starting wages will increase by $2 per hour, holding race constant.

In this example, our independent variable is *months of job training*, our dependent variable is *starting wage*, and *race* is a control variable. If this is the extent of the interpretation of the results, we cannot know if the authors are overlooking an inequitable outcome, but we would be rightly suspicious that this is the case. If participants' race was used as a control variable in the model they have presented, it was likely a statistically significant control variable (or why else include it in the final model?). It's possible that the difference between racial groups was negligible or quite substantive—we don't know. When using race or other characteristics as control variables, then, it is essential to explicitly describe the relationship between race and the dependent variable. We should never mindlessly include demographic characteristics as control variables just because that's what everyone does without bothering to interpret the impact of those control variables on our findings.

(5) *Do not assume white men as "normal" when using dummy variables*. If this is the first time you've encountered the term *dummy variable*, you may think I am about to caution against using the term *dummy*. Nope. That is just the jargon used to describe a certain type of dichotomous variable. If we had a regular, non-dummy variable for race, using the U.S. Census categories, our data for three survey respondents might look like this:

Name	**Race**
Ed	White
Margaret	Black or African American
Alleen	Asian American

There, we have a variable titled *Race* with the values *White*, *Black or African American*, *Asian American*, plus two others that are not represented in our data, *American Indian/Alaska Native*, and *Native Hawaiian/Pacific Islander*.

If we use *dummy coding* for our race variable, those same three survey respondents' data would like this:

Name	**White**	**Black or African American**	**Asian American**	**American Indian/ Alaska Native**	**Native Hawaiian/ Pacific Islander**
Ed	1	0	0	0	0
Margaret	0	1	0	0	0
Alleen	0	0	1	0	0

Now, we have five dummy variables, one for each race category. Each of the variables can take on the values of zero (meaning, basically, *no*) or one (meaning *yes*). This approach to organizing our data has the benefit of transforming the nominal-level data to ratio-level, which gives us many more options for quantitative analysis. Dummy variables are commonly used in regression analysis. When dummy variables are used as independent variables in regression analysis, one of the dummy variables is omitted from the analysis and becomes the reference category. Here is an example of such a regression model using abbreviated names for the race dummy variables above and a hypothetical index of attitude toward entrepreneurship:

$$\textit{Predicted Entrepreneurship Attitude} = \beta_0 + \beta_1 \textit{*Black} + \beta_2 \textit{*Asian} + \beta_3 \textit{*AIAN} + \beta_4 \textit{*NHPI}$$

Note that the *White* dummy variable is not included in the model; it serves as the reference category. This is fine; there is no one right way to select the reference category, and mathematically, it doesn't matter. Statistical software packages might select the reference group alphabetically, or we might select the category with the most cases. Sometimes, though, we select the reference category because it is considered normal or typical. If we dummy coded a COVID status variable, for example, we could have dummy variables for people who have never had COVID, people who have COVID, and people who have recovered from COVID. In this example, it would be reasonable to use *never had COVID* as our reference category because that is "normal." Here is where we must be careful with dummy variables (and if you are new to dummy variables or regression analysis, this is the important point): In presenting our findings, we must be careful not to treat different demographic groups as "normal." In the model above, we should reconsider this type of presentation of results:

Race	**Predicted entrepreneurship index score**
Black or African American	-2
Asian American	-1
American Indian/Alaska Native	+4
Native Hawaiian/Pacific Islander	+3
Reference group: White respondents	

That presentation implies that white respondents should be considered the norm—the standard to which other groups are compared. Instead, we could present, say, mean values for each group and highlight more meaningful comparisons among them.

(6) *Involve stakeholders in planning and conducting research* and (7) *examine your own biases*. I am offering the sixth and seventh recommendations together because they are closely related. In research about any group of people, it is a good idea to consult with members of that group in planning or, even better, conducting the research. I, a white, middle-aged man, have found this to be essential to learning about the attitudes of young, mostly African-American and Hispanic, people toward sex education programs in middle and high schools. By asking representatives of this group for feedback on survey items and plans for administering surveys, I was able to dodge potential misunderstandings and resistance to their peers' participation that I otherwise would not have anticipated. This is due in no small part to my own biases. I think about the world in a certain way that is shaped by my own experiences, and this will affect concrete research methods choices, like how I word questions that I ask research participants, how I invite people to participate in research, and how I go about collecting the data. It is important for me to reflect on how my own biases may influence such choices and perhaps to read about others' perspectives, but self-reflection and reading can only go so far. Inviting others to provide their ideas about research plans and engaging with diverse research collaborators are invaluable when I am conducting research about—or, put better, hoping to *learn from*—people who have had different life experiences than me.

(8) *Honor the humanity of research participants*. We could surely extend this list much further, but instead of a long list of tips, I will conclude with this guiding principle that should be foundational to all research about people, repeated from what I've written elsewhere about research ethics: Our research participants are not merely "subjects," they are neither data points nor ID numbers, they cannot be fully known by the values we assign to variables for them, and they are not individual representatives of the generalizations we hope to derive from our research (see Appendix F on this last point). The people who participate in research are individuals of inestimable worth and dignity, and they should be respected accordingly.

APPENDIX F: ECOLOGICAL FALLACY

Social science researchers often study groups of cases, especially groups of people. To draw warranted conclusions from such research, we must be very clear about whether we are drawing conclusions about groups or individuals. A common error is to attribute group-level characteristics to individuals; this error is the *ecological fallacy*. A researcher could succumb to the ecological fallacy by erroneously assigning a group characteristic—*most people like Star Wars*—to an individual—*Sally likes Star Wars*. That kind of ecological fallacy is easy enough to spot. Trickier to spot is the ecological fallacy of assuming relationships observed at the group level also describe relationships at the individual level. We can fall into this trap when we forget that group summary statistics, like the group average, can hide a lot of variation within groups.

For example, imagine we conducted a survey of volunteers at the Downtown Food Bank, Midtown Food Bank, and Uptown Food Bank, asking them how many hours they volunteer per month and whether they consider themselves to be generally happy. We want to know if there is an association between the amount of time spent volunteering and volunteers' happiness. We collect the following data:

Food Bank	Monthly volunteer hours	Generally very happy?
Downtown	21	No
Downtown	41	No
Downtown	56	Yes
Downtown	60	Yes
Downtown	48	Yes
Downtown	18	No
Downtown	36	No
Midtown	14	Yes
Midtown	12	Yes
Midtown	8	No
Midtown	10	No
Midtown	38	Yes
Midtown	46	Yes
Midtown	12	No
Uptown	6	Yes
Uptown	7	Yes
Uptown	12	Yes
Uptown	15	Yes
Uptown	1	No
Uptown	5	No
Uptown	24	Yes

We can summarize our data at the group level like this:

Food bank	Average monthly hours per volunteer	Generally very happy
Downtown	40	43%
Midtown	20	57%
Uptown	10	71%

When our unit of analysis is the food bank, we see a negative association between the average monthly hours per volunteer and a food bank's percentage of generally happy volunteers. We may be tempted to apply this finding at the individual level, concluding that people who volunteer more are less happy. What happens, though, when we conduct our analysis at the level of the individual? Let's compare the average hours worked by generally happy volunteers to the average hours worked by their less happy peers—so, still looking for a relationship between the independent and dependent variables, but this time without first grouping our cases:

Monthly volunteer hours

Generally very happy	**Not generally very happy**
28	17

Now, we see a positive association between time spent volunteering and general happiness; volunteers who describe themselves as generally happy volunteer, on average, more per month than everyone else. That's the opposite conclusion we had reached before! The difference? Before, we applied a finding about the relationship between two variables at the group level to individuals—we posited an ecological fallacy.

Social researchers are drawn to making group comparisons because they often reveal interesting patterns in our social world. We all base a lot of our own self-identities in our group memberships, so it can be easy to embrace results that confirm our biases about other groups or confirm our own positive self-perceptions, even if the results reflect an ecological fallacy.

APPENDIX G: RESEARCH METHODS GLOSSARY

This glossary provides definitions for the research methods jargon found in this book and for some other terms you might encounter as you learn more about research methods.

Accuracy, level of (in sampling): The breadth of the interval in which parameters can be estimated using statistics with a given level of confidence

Administrative data: Data collected in the course of implementing a policy or program or operating an organization

Alternative hypothesis: See *hypothesis testing*

Analytic generalizability: The extent to which a theory applies ("generalizes") to a given case; demonstrating analytic generalizability is held by some researchers as a goal for qualitative research

Antecedent variable: An independent variable that causes changes in the key independent variable, which, in turn, causes change in the dependent variable

Association: A probabilistic relationship between two or more variables

Axial coding: Organizing the themes that emerge from open coding, frequently by combining them into general themes subdivided into more specific themes and identifying additional relationships among codes, resulting in an organized set of codes that can be used in subsequent analysis of qualitative data

Bias: The systematic distortion of findings due to a shortcoming of the research design

Case study comparison research design: Research design in which multiple case studies are conducted and compared

Case study research design: Systematic study of a complex case (such as an event, a program, a policy) that is in-depth, holistic, using multiple data sources/methods/collection techniques

Case: An object of systematic observations; an entity to which we assign values for variables

Census: (1) A sample comprised of the entire population; (2) a study in which the sample is comprised of the entire population

Chunking: Identifying short segments of meaningful qualitative data to be coded and analyzed

Closed-ended question: A survey or interview question that requires respondents to select from a set of predetermined responses

Cluster sampling: A probability sampling design in which successively narrower aggregates of cases are selected before ultimately selecting cases for inclusion in the sample

Coding: See axial coding, open coding, selective coding

Concept: An abstraction derived from what many instances of it have in common

Concurrent validity: A type of criterion validity describing the extent to which a variable (or set of variables intended to operationalize a single concept) relates to another variable measured at the same time as would be expected if the variable accurately measures what it is intended to measure

Confidence, level of (in sampling): The certainty, expressed as a percentage, with which parameters can be estimated using statistics with a given level of accuracy; the percentage of times an estimated parameter would be expected to be within a given range (the level of accuracy) if calculated using data collected from a large number of hypothetical samples

Confidence interval: The range of values we estimate a population parameter to fall in at a given level of confidence

Content validity: An aspect of operational validity describing the extent to which the operationalization of an abstract concept measures the full breadth of meaning connoted by the concept

Control variable: A variable that might threaten nonspuriousness when examining the causal relationship between an independent variable and dependent variable; control variables are plausibly related to both the independent and dependent variables and could thus explain an observed association between them; in an experiment or quasi-experiment, control variables are those variables held constant so that they cannot affect the dependent variable while the independent variable is manipulated

Convenience sampling: A nonprobability sampling design in which cases are selected because they are convenient for the researcher

Conversational interviews: Interview conducted following a very flexible protocol outlining general themes but permitting the interview to evolve like a natural conversation between the researcher and respondent

Criterion validity: An aspect of operational validity describing the extent to which a variable (or set of variables intended to operationalize a single concept) is associated with another variable as would be expected if the variable accurately measures what it is intended to measure

Cross sectional research design: A formal research design in which data are collected in one "wave" of data collection, with data analysis making no distinction among data collected at different times

Data analysis: Systematically finding patterns in data

Dependent variable: A variable with values that are dependent on the values of another variable; in a cause-and-effect relationship, the variable representing the effect

Descriptive data analysis: Quantitative data analysis that summarizes characteristics of the sample

Discriminate validity: An aspect of operational validity describing the extent to which the operationalization of an abstract concept discriminates between the target concept and other concepts

Disproportionate stratified sampling: A probability sampling design in which the proportions of cases in the population demonstrating known characteristics are intentionally and strategically different for the cases in the sample, usually to permit comparisons among subsets of the sample that may otherwise have had too few cases

Dissemination: To share the results of a study and how it was conducted widely, usually by publication

Double-barreled question: A question, such as in an interview or survey, that is actually asking two questions at once

Dummy variables and **dummy coding**: A dummy variable is a variable that takes on two values: one (meaning, basically, *yes*) and zero (meaning *no*). Dummy coding is the process of transforming a single categorical variable into a series of dummy variables, with each value of the original categorical variable transformed into its own dummy variable. For example, the variable *student classification* with the values *freshman*, *sophomore*, *junior*, and *senior*, can be transformed into four dummy variables, *freshman*, *sophomore*, *junior*, and *senior*, each taking on the values of one or zero. Dummy coding a categorical variable thus yields a series of ratio- level variables, enabling a much wider range of quantitative analysis options.

Ecological fallacy: A research finding made in error by mistakenly applying what has been learned about groups of cases to individual cases

Effect size: A quantitative measure of the magnitude of a statistical relationship **Empirical research**: Generating knowledge based on systematic observations **Empirical**: Based on systematic observation

Empiricism: The stance that the only things that are "real" and therefore matter are those things that can be directly observed; not to be confused with *empirical*

Experimental research design: A formal research design in which cases are randomly assigned to at least one experimental group and one control group with the researcher determining the values of the independent variables that will be assigned to each group and the dependent variable measured after (and usually before as well) manipulation of the independent variable

External validity: The generalizability of claims generated by empirical research beyond cases directly observed

Face validity: An aspect of operational validity describing the extent to which a variable (or set of variables intended to operationalize a single concept) appears to measure what it is intended to measure

Fact-value dichotomy: The naïve view that *fact* and *value* are always wholly distinct categories

Focus group:A group of individuals who share something in common of relevance to the research project who are interviewed together and encouraged to interact to allow themes to emerge from the group discourse

Generalize: To make claims beyond what can be claimed based on direct observation, such as making claims about an entire population based on observations of a sample of the population

Hawthorne effect: Bias resulting from changes in research participants' behavior effected by their awareness of being observed

Hypothesis: A statement describing the expected relationship between two or more variables

Hypothesis testing: A method used in inferential statistics wherein the statistical relationships observed in sample data are compared to a hypothetical distribution of data in which there is no analogous relationship to generate an estimate of how likely or unlikely the observed relationship is; the observed relationship being tested is stated as the *alternative hypothesis*, which is compared to the statement of no relationship, the *null hypothesis*

Independent variable: A variable with values that, at least in part, determine values of another variable; in a cause-and-effect relationship, the variable representing the cause

Inferential data analysis: Quantitative data analysis that uses statistics to estimate parameters

Informed consent: An individual's formal agreement to participate in a study after receiving information about the study's risks and benefits, assurances that participation is voluntary, what participation will entail, confidentiality safeguards, and whom to contact if they have questions or concerns about the study

Institutional Review Board: A committee responsible for ensuring compliance with ethical standards for conducting research at an institution, such as a university

Internal validity: The truth of causal claims inferred from empirical research

Interval scale of measurement: Describes a variable with numeric values but no natural zero

Intervening variable: An independent variable that itself is affected by the key independent variable and then, in turn, causes change in the dependent variable

Interview protocol: The set of instructions and questions used to guide interviews

Latent variable: A variable that cannot be directly observed, such as an abstract concept, attitude, or private behavior

Literature review: (1) The process of finding and learning from previous research as one of the early steps in the research process; (2) a paper that summarizes, structures, and evaluates the existing body of knowledge addressing a research question; (3) a section of a larger research report that summarizes, structures, and evaluates the existing body of knowledge being addressed by the research and locates the research being reported in that larger body of knowledge

Logic model: A diagram depicting the way a program is intended to work, including its inputs, activities, outputs, and outcomes

Manifest variable: A variable that can be observed and is thought to indicate the values of latent variable

Memoing: Writing notes to document the qualitative researchers' thought processes associated with every step of qualitative research and their evolving ideas about what is being learned during the course of data analysis

Meta-analysis: A method of synthesizing previous research using statistical techniques that combine the results from multiple separate studies; the results of research using this method

Mixed methods research: Research using both qualitative and quantitative data

Natural experiment: A quasi-experimental design that capitalizes on "naturally" occurring variation in the independent variable

Nominal scale of measurement: Describes a variable with categorical values that have no inherent order

Nonparametric data analysis: Analysis of quantitative data using statistical techniques suitable because the data do not have an underlying normal distribution, homogeneous variance, and independent error terms

Nonprobability sampling design: A strategy for selecting a sample in which the probability of cases being selected is either unknown or not considered when selecting cases for inclusion in the sample, with sample selection made for some other reason (see *convenience sampling*, *purposive sampling*, *quota sampling*, and *snowball sampling*)

Nonspurious: Not attributable to any other factor

Null hypothesis: See *hypothesis testing*

Open coding: Assigning labels/descriptors/tags to “chunks” of qualitative data that note the data’s significance for addressing the research question; a first step in identifying important themes that emerge from qualitative data

Open-ended question: A survey or interview question without any predetermined responses

Operational validity: The extent to which a variable (or set of variables intended to operationalize a single concept) accurately and thoroughly measures what it is intended to measure

Operationalize: To describe how observations will be made so that values can be assigned to variables for cases

Ordinal scale of measurement: Describes a variable with categorical values that have an inherent order

Panel research design: A formal research design in which data are collected at different points across time from the same sample

Parameter: A quantified summary characteristic of a population

Parametric data analysis: Analysis of quantitative data using statistical techniques suitable only because the data have an underlying normal distribution, homogeneous variance, and independent error terms

Peer review: The process of having a research report (or other form of scholarship) reviewed by scholars in the field, usually as a prerequisite for publication

Plagiarism: The written misrepresentation of someone else’s words or ideas as one’s own

Point estimate: A statistic calculated from sample data used to estimate the population parameter; usually referred to in distinction to the *confidence interval*

Policy model: An explanation of how a policy is supposed to work, including its inputs, how it is intended to be implemented, its intended outcomes, and the assumptions that undergird the intended change process

Population: Total set of cases of interest; all cases to which the research is intended to apply

Predictive validity: A type of criterion validity describing the extent to which a variable (or set of variables intended to operationalize a single concept) predicts future change in another variable as would be expected if the variable accurately measures what it is intended to measure

Probability sampling design: A strategy for selecting a sample in which every case in the population has a known (or knowable) nonzero probability of being included in the sample

Proportionate stratified sampling: A probability sampling design in which the proportions of cases in the population demonstrating known characteristics are replicated in the sample

Purposive sampling: A nonprobability sampling design in which cases are selected because they are of interest, typical, or atypical as suits the purposes of the research

Qualitative data: Textual data

Quantitative data: Numeric data

Quasi-experimental research design: A formal research design similar to experimental research design but with assignment to experimental and comparison groups made in a nonrandom fashion

Quota sampling: A nonprobability sampling design in which cases are selected as in convenience sampling but such that the sample demonstrates desired proportions of characteristics, either to replicate known population characteristics or permit comparisons of subsets of the sample

Ratio scale of measurement: Describes a variable with numeric values and a natural zero

Reliability: The extent to which hypothetical repeated measures of variables would generate the same values for the same cases

Research design: 1) Generally, a description of the entire research process; 2) more narrowly, the formal research design used to structure the research, including cross-sectional, time series, panel, experimental, quasi-experimental, and case study research designs

Response set bias: Bias resulting from a response set that leads respondents to select responses other than more accurate responses

Response set: The set of responses that respondents may select from when answering a closed-ended question

Sample: Subset of population used to learn about the population; the cases which are observed **Sampling error**: The difference between a statistic and its corresponding parameter **Sampling frame**: List of cases from which a sample is selected

Secondary data: Data collected by someone other than the researcher, usually without having anticipated how the data would ultimately be used by the researcher

Selective coding: Assigning a set of codes (such as a system of codes developed through axial coding) to "chunks" of qualitative data

Semi-structured interviews: Interviews conducted following an interview protocol that specifies questions and potential follow-up questions but permitting flexibility in the order and specific wording of questions

Simple random sampling: A probability sampling design in which every case in the population has an equal probability of being selected for inclusion in the sample

Snowball sampling: A nonprobability sampling design in which one case is selected for the sample, which then leads the researcher to another case for inclusion in the sample, then another case, and so on (also called *network sampling* when cases are people)

Social desirability bias: The tendency of interviewees to provide responses they think are more socially acceptable than accurate responses

Standardized interview: Interviews conducted following an interview protocol requiring identical wording and question order for all respondents

Statistic: A quantified summary characteristic of a sample

Systematic sampling: A probability sampling design in which every *k*th case in the sampling frame is selected for inclusion in the sample; if there is a discrete (as opposed to hypothetically infinite) sampling frame, *k* equals the number of cases in the population divided by the number of cases desired to be in the sample

Theory: A set of concepts and relationships among those concepts posited in a formal statement to describe or explain the phenomenon of interest

Time series research design: A formal research design in which data are collected at different points across time from independent samples

Unit of analysis: The entity—the whom or what—that is being studied; the entity for which observations are being recorded in a study

Validity: Truthfulness of claims made based on research; see *operational validity*, *face validity*, *content validity*, *discriminate validity*, *criterion validity*, *concurrent validity*, *predictive validity*, *internal validity*, *external validity*

Variable: Logical groupings of attributes; the category to which these attributes belong; a factor/quality/condition that can take on more than one value/state