LEARNING RESEARCH METHODS

A SOCIAL STUDIO APPROACH

Edition 2.1

Mark Gunty
University of Notre Dame

Project One: Interview Survey

SKILL DEVELOPMENT FOCAL TASKS

- Writing survey questions
- Conducting face-to-face interviews
- Recording responses to an interview
- Designing a record observation sheet
- Writing simple hypotheses
- Identifying appropriate control variables

SURVEY OVERVIEW, PART 1

It seems that surveys are everywhere, in the form of telemarketing phone surveys, public opinion polls, mail-in surveys, student government sponsored questionnaires, and feedback surveys after experiencing a service. With all these surveys, one could get the impression that it is very easy to put together a survey. In fact, it is not.

Your first project is to design and conduct a face-to-face interview survey. Later, for the fourth project, you will work on another survey which will be self-administered. The basic difference between interviews and self-administered surveys is that the former consists of an interaction during which the researcher records the responses and comments of the respondent, while the latter consists of having the respondent provide answers directly through some kind of instrument. It's kind of like the difference between a conversation and an exchange of letters.

This first project, therefore, will focus on your interview skills. Learning to ask questions effectively, however, is part of any survey, so that skill will be further developed in Project Four. Although it is not a necessary distinction between interviews and self-administered surveys, it so happens that for this first project you will learn the basics of drawing a nonprobability sample and your other survey project will require a probability sample. That difference will be explained later.

It is easy to understand why surveys are so popular, especially if done well. More than anything, surveys enable the researcher to generalize to large populations like no other research strategy. If the research objective involves understanding trends in society or getting the pulse of a particular group, survey is frequently the tool of choice. Second, by definition surveys rely on self-report, so if the concepts in a study include thoughts, feelings, opinions, attitudes or beliefs, surveys can operationalize them quite nicely. Third, a good survey instrument can measure dozens—even hundreds—of variables, making it possible to examine multiple relationships with a single data collection effort. Finally, for the amount of

information generated, surveys are relatively inexpensive, rivaled only by archival analysis in terms of cost efficiency.

As versatile as they are, surveys cannot do everything. Because of their reliance on self-report, even good scales (which are multiple-item indices of complex concepts like self-esteem and prejudice) will be vulnerable to criticism about measurement validity. Of course, single-item measures can be even more troublesome. The connection between how people describe themselves and how they actually behave varies, leaving the researcher uncertain about the usefulness of surveys for examining behaviors. The underlying problem is the impossibility of collecting information without the survey respondent knowing that he or she is being observed (see the sidebar on pp 12-13 on the criticisms of social science that are related to the consciousness of the observed). Surveys tend to generate more reactivity problems than other strategies. Additionally, establishing causality with survey data requires longitudinal designs which can offset the advantages of cost efficiency and generalizability. Finally, surveys depend heavily on verbal interaction (either written or spoken), making them less useful with very young children or others for whom language is difficult.

We usually think of a set of questions as a survey, but actually three things make up the kind of research called a survey: the interview or questionnaire (also called the instrument), the sample, and the administration strategy. *In other words, you need to ask the **right questions** of the **right people** in the **right way**.* These three pieces are all essential. Getting any one of them wrong will ruin the entire project. To repeat, it is not easy to do a *good* survey. Most of the features of survey design will be covered in the second survey project. This time around, the focus will be limited to a few basics (see Focal Tasks above). We will start with the last of those components and deal with the way the survey is administered. Administration includes three decisions: (1) the data collection mode, (2) time design, and (3) schedule.

Data Collection Modes

As mentioned above, there are two basic strategies for making the observations involved in survey research. On the one hand, you can have a researcher listen to the **respondent** (the term for a participant in a survey) and record the information. In other words, you conduct **interviews**. On the other hand, you can give respondents the instrument with a means for them to record their own responses. This approach makes the survey **self-administered**. Interviews tend to build relationships, while self-administration allows for some privacy. Choosing the appropriate data collection mode depends on a number of factors. You need to have a very good sense of your research purposes, resources and constraints, and population. The advantages and disadvantages of self-administered surveys will be reviewed in Project Four. Here we will concentrate on interviews.

Interviews can be conducted in person face-to-face, video face-to-face or by phone. **In-person interviews** require somewhat elaborate arrangements because both interviewer and interviewee need to be in the same place. The interviewer will usually have to make the

primary effort to conform his or her schedule to the interviewee's and to make the rendezvous happen. For some types of **phone interviewing**, the arrangements can be fairly simple. If the calls are made to homes, evenings are usually the best contact time. The call itself is the time to determine whether the respondent is available. That is because, unlike face-to-face, there is very little cost or time involved in trying again. For phone interviews with people at their place of work, it will typically take at least one phone call to explain the purpose of the interview and to arrange a good time to conduct the interview, like setting a meeting on the respondent's calendar. Then the interviewer should call back promptly at the appointed time. We follow this procedure on the assumption that most people at work will not be available for a significant chunk of time when the researcher makes the initial call.

Time Order Designs

Before describing the time design options, let's think for a minute about causality, which you will learn in greater detail later. Establishing time order is one of the three necessary criteria for determining causality, because causes must precede effects. A person completes a survey at a single point in time, so the measures derived from the survey lack time order. As you will quickly learn, there are some ways to address this limitation, but in general the lack of time order presents a serious obstacle to establishing causality using surveys.

When information for survey research is collected all at one time, it is called a **cross-sectional design**. Often, though, that is good enough, because respondents can give you information about things that happened earlier in their lives and you can approximate time order. For example, it is common to survey students as they enter college about both their experiences in high school and their reasons for choosing a particular college (which are part of their past), while also asking questions about their expectations for their college experience (which is part of their present). This approach can only be used when the information needed about the past is easy to recall accurately and not complicated in nature. Of course, your causal claims will need to be put in context, but they will not be unfounded. Alternatively, cross-sectional studies work well if the research objective is to describe population characteristics at a single point in time. That is the strength of public opinion polls.

You have two options if you decide to collect survey information at more than one point in time. One type of **longitudinal design** is called a **trend study**. At several points in time, you administer questionnaires, each time to different people. Trend studies are used to trace historical trends. The research focuses on general, aggregate-level change within populations. This option is also referred to as the **repeated cross-sectional design**.

Your other option is a **panel study** (also called a **fixed sample**), in which you administer surveys repeatedly to the same people. The term "panel" refers to a sample that is used more than once, surveying a select group of people at one point, then locating them and surveying the same people again some other time (and perhaps additional times as well).

The use of the panel clearly distinguishes this longitudinal design from the trend study, which draws a new sample each time the survey is administered. A panel can be defined in a number of ways, such as a cohort that started something together (e.g., entering school or a profession, or born in the same year), otherwise dissociated people who experienced something together (e.g., military service, a natural disaster, or changes in leadership in a religious congregation), or randomly selected people who are followed over time. With this design, obviously, time order is apparent by the sequence of time points at which various measures are taken. Panel studies, while being very useful designs for conducting explanatory research for large populations, have a certain vulnerability to **attrition**, which happens when members of your initial panel cannot be located or who otherwise choose not to participate in subsequent rounds of data collection.

The problem that remains, even with the best of panel studies and certainly with trend studies, is a sticky one: having observed change from one time point to another, which of the many things that happened between those two time points can be identified as the cause? One of the reasons that surveys measure so many variables is because they have to. Survey data is subject to many plausible explanations, necessitating the application of sophisticated techniques to control for dozens of factors and to test for association along multiple parallel causal paths.

Scheduling

Scheduling of a survey differs from time designs and refers simply to plans for doing which things at what time. Once you have a draft of your instrument, certain things have to be done in order:

1. Pilot testing of survey and/or training of interviewers
2. First wave of contacting the sample
3. Second wave of contacting the sample
4. Subsequent waves of contacting the sample

For this project you will get your first taste of the importance of **pilot testing**, the stage of research when you test how sound are your design decisions. Because it is a longitudinal panel study, the project has considerable flexibility regarding when you collect each wave of data, but you will nonetheless have to make decisions about timing. What days of the week are best? What time of day? How long should lapse between interviews? Is there anything special you have to do for the first or last interviews? These are all part of scheduling. Other aspects of scheduling will be incorporated into the second survey project.

ETHICS

Surveys seldom expose participants to dangerous risks, so the main ethical concerns in relation to protection of human subjects are **non-coercion** and **protection of privacy**. To guard against people feeling forced to participate, or even the appearance of coercion of any

kind, the researchers will rely on the principle of **informed consent**. Informing potential respondents does not require that you reveal the entire purpose of the survey; in fact, that might be counterproductive if it increased reactivity. It is usually sufficient to communicate the general topic of the study, the sponsoring or funding organization, the affiliations of the researchers (their college, university, think tank, or company), and, occasionally, the method by which the individual was selected for participation. When the participants are chosen because of their membership in an organization or school, the cover letter or introduction must inform them that their decision to participate or not does not affect their standing in that organization. On the consent side, an individual's willingness to complete a survey or to submit to an interview is considered consent to participate. The researcher has an obligation, however, to make it clear that the participant can withdraw consent at any time.

The protection of privacy for surveys means that the data you collect should be secured against hacking or any other form of unauthorized access. This includes all paper copies of questionnaires, interview notes, and electronic databases. Further you will take measures to make sure that the identity of participants is not released. My mantra is, "No individually identifiable information will be released in any form." You must adhere to this dictum rigorously, going so far as to anticipate that releasing some descriptive information along with survey responses amounts to revealing an individual's identity. For example, simply reporting survey results by gender and ethnicity might compromise the privacy of those individuals who are one of two or three people in the sample with a particular combination of those traits (like being one of just three male Latinos).

If you collect the data such that you gather names, you must pledge **confidentiality**. That means that you will not reveal identities, but you, the researcher, do know them. If you decide to conduct the survey such that you do not ask respondents to identify themselves, your survey is **anonymous**. Anonymity is not ethically necessary. Indeed, it can be methodologically challenging if not impossible. If you are collecting longitudinal data (see below), you will need some way to link a respondent's inputs from one time point with his or her inputs at the other time points. There are strategies to do so that leave the data less vulnerable than others. You can, for example, assign study participants a code. In one document you have the key for which code belongs to which person, but on all of the questionnaires only the code is recorded. Then, after the last wave of data collection and linking all of the waves of data together, you can destroy the key. Or you can use a system of initials if it is a small project. In this interview project, that is the system you will use.

In addition to your ethical obligations regarding the protection of human subjects, you also have an ethical obligation to conduct the research and to present the results of that research honestly. It may seem self-evident, but it is worth mentioning that you should never fabricate or alter the data. In the discussion of results, refrain from tainting your conclusions with unfounded speculation.

CRITICISMS OF SOCIAL SCIENCE. BOX 1

Despite the value of social science research, it has come on the scene relatively recently compared to the physical and life sciences. Indeed, it is easy to think of reasons why social science should not be conducted, or at least not called science. Throughout this workbook, we will consider some of those criticisms. They draw attention to features of our work that deserve particular attention. In reviewing these criticisms, we shall distinguish between evaluating them in relation to the social aspect of social science and comparing them to similar criticisms of the physical sciences.

The first criticism is that the consciousness of the objects of study interferes with the observation process. This problem is referred to as **reactivity**. People who know they are being observed tend to change their behavior in a variety of ways. Physical scientists usually do not have this problem, although animal researchers sometimes do. If, when we try to study humans, they do not behave the way they would if we were not studying them, what is the point? It is true that we cannot observe social reality without changing it, and this problem presents several serious challenges. When and where possible, we have to learn to observe things in ways that leave our subjects unaware of our observation. In the language of research, we should be *unobtrusive*. In many field research projects, researchers may discover ways to be totally unobtrusive and blend in smoothly with the setting. They may also choose to reveal little to nothing about their research intentions so that, while people know someone is observing, it doesn't feel like anything different from day to day life.

When unobtrusiveness is not possible, which is nearly always the case with surveys, we need to be aware of the effects of observation and take them into account when analyzing data. In survey research, obtrusiveness means that respondents will tend to misrepresent themselves in order to project an image that they think is, on the one hand, what society generally expects of people (**socially desirable**), and, on the other hand, what respondents believe the researcher expects of them (**researcher effects**). Note, however, that some respondents try to do the opposite of what they think is expected, but, either way, reactivity is a problem because the information gathered is not accurate.

MEASUREMENT

Measurement is the process of capturing values that reflect the concepts you want to study. It answers the question of how much or what kind, so keep in mind that measurement is not restricted to quantities.

Recall that in Exercise 1 you had some experience with conceptualization (defining terms) and operationalization (specifying procedures) and that those elements were the first and last steps in the **CVIO model**. Now you are introduced to the middle two steps:

1. Define the Concept
2. Identify Variables
3. Select Indicators
4. Operationalize the Measure

Once you have a concept defined, you think of characteristics, traits, and other manifestations of the concept. Most concepts can be converted to variables in many ways. For example, religiousness, defined as engagement in the institutional practice of religion, would be reflected in any of the following variables:

- Frequency of participation in religiously focused activities
- Extent to which one identifies with one's religious affiliation
- Amount of religious artifacts in one's living space
- Frequency of private prayer
- Depth of participation in the ministries and leadership of one's place of worship

Note that these phrases all refer to characteristics and behaviors that vary from one person to the next. Their wording is intended to capture that variation ("frequency of," "extent to which," etc. in these examples and "type of," "kind of," and so on for categorical variables). Throughout the full range of projects in this course, you will have repeated opportunities to identify variables.

The third step in the CVIO model is to select indicators. Variables are characteristics, and indicators are observable manifestations of the variables. There are four kinds of **indicators**: self-report, other-report, behavioral observations, and artifact observations. All four will be explained eventually, but for this project we will concentrate on **self-report** which consists of the subject telling the researcher something about himself or herself. As mentioned above, self-report is at the heart of survey research. Notice that self-report is observable (available to the senses), either by hearing the response or seeing how the respondent marked the questionnaire.

Self-report is simply a type of indicator. The third step in the CVIO model would more properly be described as something like, "Self report of the number of times in the last month that the respondent attended a non-worship church-related activity." Furthermore, it is common to operationalize a concept with multiple indicators. To continue with the religiousness example, suppose you had designed a questionnaire that asked how frequently in the last month someone attended any of the following church-sponsored events: (a) social gatherings, (b) planning or committee meetings, (c) ministry training, and (d) social service activities. A person gives the answers 2, 1, 0, 3, respectively to the four types of events. In a sense, you have four self-reports and you operationalize the variable by taking the sum of the answers. This serves as a basic illustration of indicators and will suffice for this project; keep in mind that you will learn much more about operationalization in the coming weeks.

Measures come in a wide variety of forms. One aspect of a measure is the **level of measurement**. This project will merely introduce you to the notion of levels because it is hard to create survey questions without consciously applying different levels. In the next

project you will get a more detailed explanation. There are three levels used in social science.[1]

- **Nominal or categorical measures** represent categories or types. Something could be in one and only one of the categories, but the categories have no mathematical relation to each other, as in religious affiliation (Muslim, Christian, Jewish, Buddhist, etc.). When a nominal measure has only two categories (e.g., male or female), it is a **dichotomous variable**. When the dichotomy is conceived as either the presence or absence of something (such as "Lives with father" and "Does not live with father"), it is a **dummy variable**.
- **Ordinal measures** represent an amount of something, but in a way that can only be described in terms of more or less than the other values, not how much more or less. One kind includes things like birth order or grade in school. Another kind of ordinal measure is the value on a scale, for example, from very satisfied to very dissatisfied or from strongly agree to strongly disagree. We can say that a person who is very satisfied shows more satisfaction than a person who is satisfied, but we cannot say *how much more* satisfaction.
- **Ratio measures** have a meaningful zero, fixed intervals, and a range from high to low, making it possible to multiply and divide them. Ratio measures abound: age, income in dollars per year, number of employees, number of times a word appears in an article, and so on.

Because it is highly likely that the design of the interview survey will include at least one categorical measure, know that a requirement for questionnaire items with categorical responses is that the set of options should be **exhaustive and mutually exclusive**. That means that (a) all possible categories are listed (exhaustive) and (b) a response will fit into one and only one category (mutually exclusive). Continuing with our religious affiliation example, do not list both Lutheran and Protestant, since those categories overlap. Since a truly exhaustive list for some categories would be impractically long (e.g., occupations or reasons for choosing a college), it is best to list the categories likely to fit most of your cases, then include the categories "None" and "Other." You may or may not want a fill-in space by that category (e.g., "Other (please specify):_____").

Finally, one other special measurement application is something called the **change variable**. The nature of a change measure is a computation of the difference between observations at two points in time. That is, we want to see the extent to which and direction in which something changed. For example, we might look at frequency of attendance at religious services by married couples before and after having children. Our concern is not so much whether they are attending weekly or monthly either before or after children, but whether they do so more or less frequently after children. A very important type of change

[1] Many texts present four levels of measurement. This text leaves out the interval level, which, for all practical purposes, does not exist in the social sciences. Scales, often presented as interval measures, are more accurately described as ordinal measures with many values.

variable is academic performance. If you want to study the effects of some special instructional method, you need to know more than how students performed on standardized tests at the end of a time period in which the special method was implemented. You need to know how their test scores changed from the beginning of the time period to the end.

Change measures can be computed as the difference between the score on the "after" observation and the "before" observation (e.g., 89% correct on the posttest minus 83% on the pretest yields a change score of +6%). Alternately, the change score can be computed as a percentage change. The average weekly number of times a person calls home in the four weeks of the semester is 6; in the last four weeks of the semester it is 4.5, yielding a change score of negative 25% (new minus old divided by old). Note that a change score can be positive or negative, indicating an increase or decrease over time.

INSTRUMENT CONSTRUCTION

Because we will try to keep the interview for this project very short, here we only review a few general principles about instrument construction. "Instrument" is the generic term for the means of collecting all of the participants' information; in self-administered surveys the instrument is referred to as the **questionnaire**, whereas in interviews it is called the **interview schedule**. In any case, the terminology is employed imprecisely by most people, so getting the *terms* right is not that important. Getting the *construction* right, on the other hand, is very important. There are four standard elements of instrument construction: Instructions, Questions, Response Categories, and Placement. Let's go over them.

Questions

First, "questions" on a survey are not necessarily questions in the strict grammatical sense. A better term is *item*, covering all of the options available on the prompt side of the prompt-and-response interaction between researcher and respondent.

The primary and unwavering criterion for a good survey question is unbiased clarity, which means that it elicits the information you want to collect, without influencing the outcome. Most of the time that requires wording that is exact, yet simple—but the bottom line is that respondents understand what you are asking for. Sometimes I have found that making a question very, very precise actually makes it harder to understand, so rigorously pilot test everything.

Be careful to avoid undefined terms, ambiguity, and vagueness. Ambiguity dilutes validity because you cannot be sure what information the respondent is giving you. For example, you may want to avoid wording such as, "How often do you get in trouble with your parents?" It would probably help to be more precise about "getting in trouble." On occasion, it is acceptable to use broad language such as, "Overall, how satisfied were you

with the services of hospital staff during your stay?" In that case, the question is intentionally broad, allowing the respondent to bundle together all of his or her impressions into an overall rating. Ambiguous wording, such as "cheating" or "hooking up," which are known to mean many different things to different people, should be avoided.

Open and Closed Response Categories

One option in the survey-maker's toolkit is the open-ended question. That is, you pose the question and let the respondent answer however they wish. The advantage to this strategy is that the response is not constrained by pre-conceived notions the researcher has about the issue, notions which define and limit the response categories provided in closed-ended questions. Additionally, in the case of interviews, open-ended questions create the opportunity for the respondent to provide information the researcher might not have thought to ask about. These advantages are offset by two disadvantages. First, open-ended questions require more work from the respondent, especially in self-administered surveys. Generally, it is easier to look down a list and select from multiple choice options than it is to write out thoughts, phrases, or whole ideas. Even in an interview (especially over the phone) it might be easier to respond "Approve" or "Disapprove" than to explain one's stance on many issues. Second, open-ended answers require more work for the researcher in terms of recording and analyzing the observation. Because of these disadvantages, open-ended questions should be used sparingly and with good purpose. In exploratory stages of research and with a cooperative sample of respondents not pressed for time, it may be useful to ask open-ended questions so that the underlying issues emerge organically, almost like they do in field research. You might use an exploratory study to determine the kinds of programs and spontaneous events in a residence hall that promote or hinder openness to diversity or sexual harassment awareness. Later, when studying a larger college population, you could convert those open-ended responses to closed-ended lists. Finally, if the question solicits small bits of information that are likely to cover a wide range of options, open-ended questions can work well. For example, it is probably easier to ask "What is your job title?" as open-ended, because job titles are usually short (easy for the respondent to write out the words) and because it would be cumbersome to list every possible job title of all respondents.

When the question is asked closed-ended (a list of response options is provided), the possibilities are quite broad. The following list is suggestive, not comprehensive, and is intended simply to generate ideas.

- Position or evaluation scales: Agree to Disagree, Excellent to Very Poor, Approve to Disapprove, Satisfied to Dissatisfied, and so on.
- Descriptive categories: ethnicity, religious affiliation, political party, highest degree attained, major, industry in which they work, job title, marital status, or personality type, just to mention a few.

- Amounts or frequencies: Frequently to Never, A Great Extent to Not at All, 20 or more hours per week to 0-1 hour per week, and so on.
- Semantic Differential. Used when you are asking respondents to give you their impression of something or someone, **semantic differential** scales might help. These scales set two opposite words on either end of a continuum and ask the respondent to show where their impression falls on the continuum. For example, you might name a public figure and have respondents indicate where on a scale of Decisive to Indecisive they would place that figure. Or the scale could be Effective-Ineffective, Charismatic-Dull, Outgoing-Shy or Trustworthy-Untrustworthy.

Placement

Placement refers to where a question falls in the questionnaire relative to the other questions. As such, placement consists of both where an item is relative to beginning, middle, or end, and what other items come before or after it. The beginning of the survey should have items that are relatively easy and non-threatening, yet engaging. The end of the survey includes items that do not involve a great deal of recall, such as simple demographics. The items at the end of a survey are the ones you are most likely to lose to item non-response, so consider whether you can live without that information.

That leaves most everything else for the middle. The object of the middle is to gradually work your way into more sensitive and complex questions. To do so, consider other factors such as keeping items in logical or chronological order if applicable. It is a good general rule not to make respondents jump around from one topic to another and back again.

SAMPLING

One of the three dimensions of validity is generalization, the extent to which findings from one set of observations can apply with confidence to a larger part of the whole. Certainly it is desirable to achieve broad generalizability, but validity is the accurate description of the amount and conditions of whatever generalizability the study warrants. Knowing how you sampled your observations from the complete set of possible observations makes that accuracy possible. Remember that observations and populations are not limited to groups of people; events, artifacts, organizations and many other things can constitute a **population**. A single one of the things in a population is an **element**.

You start your sampling strategy by defining the group to which you would like your results to apply. That is your **target population**. Having identified the target population, you must determine whether you can gain access to any listing or combination of listings that includes all elements in that population, which is called a **sampling frame**. If you have access to a sampling frame, you can and usually should use **probability sampling strategies**. Note, however, that the existence of a sampling frame does not make good sampling automatic.

Sampling frames are often not quite complete listings of the target population because of inaccuracies in contact information or lapses in record-keeping. The listing that in fact constitutes what you can sample is called the **study population.** In addition to errors in the sampling frame, other sources of difference between the target population and study population can come from practical limitations on what kind of sample you can draw. For example, your target population might be junior high students in the United States, but you are limited to states which administer state-wide math and reading tests to both fifth and eighth graders. Another source of difference between target and study populations could come from things like who happens to be home when you are canvassing a neighborhood door to door, or who is not out sick on the day you administer a survey at a school or workplace.

If no sampling frame exists or can be constructed, then you must use **nonprobability strategies**. Sampling is a multifaceted skill which will be discussed throughout all of the projects. This project presents some special limitations, so the discussion of sampling will focus on nonprobability approaches even though you could conduct this research with probability techniques.

The target population for this study is undergraduate students at your institution. A sampling frame is most certainly available because your institution keeps very close track of who is and is not considered a student. Two conditions, however, make nonprobability sampling appropriate for this project. First, you need to fit the interviews into your weekly schedule and that of your interviewees. Pulling a good representative probability sample would create a study population of students whom you do not know and whose schedules might not be at all compatible with yours. If you were a full-time researcher you would make the schedules work, but that is not necessary for this first project when your skill-building tasks are focused elsewhere.

Second, this study is designed at the descriptive level. A descriptive study falls on a continuum from explanatory to exploratory in nature. Explanatory studies derive from thorough searches of the literature, seek to contribute significantly to theory-building, and are based on specific hypothesis-testing designs. On the other end of the continuum are studies that are more exploratory in nature, making initial inquiries into unstudied populations with less concern for sweeping generalizations and more concern for testing methodologies. Descriptive studies fall closer to the exploratory side. While we hope to discover some things on a few topics of student behaviors and we might be able to hypothesize that certain relationships exist, curiosity and skill building are far more important at this stage. This explanatory-exploratory dimension is called the **level of inquiry**. One additional special type of inquiry falls closer to the explanatory end and is called evaluative. Evaluation research examines the effects of a program of some kind. It usually combines descriptive and explanatory work because a good bit of the work involves describing how the program emerged, yet most programs are designed to have specific outcomes. The program (or sub-programs) is treated as the cause and the research tests

whether it had the desired effects. Therefore, the hypothesis testing dimension of evaluation research is driven by program intention more than by theory, though the program might have been designed based on theory. The subject of evaluation research will be revisited in the Conclusion of this manual.

Let us return to the discussion of sampling. Your design will be nonprobability because you will choose participants in the study based on a **convenience sample** rather than randomness. That is, you will invite fellow students, with whom you expect to have regular contact through the semester, to take part in our study. The group from which the participants will be drawn, then, will consist of the sum of acquaintances of the students in this class. You can easily see that the group is hardly representative of the whole population of undergraduate students at your school, so no sample drawn from them can be expected to represent the whole. You can improve the representativeness by applying **quota sampling**. Using that strategy, together the students in class should plan around certain constraints, for example, equal numbers of first, second, third and fourth year students; males and females; students from the various disciplines and majors; on- and off-campus students; and so on. The value of quota sampling in this case is not so much about achieving representativeness (which would be quite a stretch), but to make sure that the questionnaire design and interview experience afford you with a wider range of cases and broader opportunities to learn.

Nonprobability sampling is not merely a convenience when a project team is short on time or resources. The fact is, some populations can only be studied with nonprobability sampling because no sampling frame could be assembled for the population, no matter how hard we try. Deviant and underground populations—such as gangs, non-incarcerated active criminals, and cults, to name just a few—are often subject to nonprobability sampling. Other populations might require nonprobability sampling as well. Examples include home-schoolers, persons with certain disabilities, vegetarians, owners of rare dog breeds, families that generate their own electricity ("living off the grid"), users of alternative medical practices, and so on. There may be some listings of people in these populations, but a complete sampling frame is often unattainable.

Most populations which require nonprobability sampling consist of people (sometimes groups), as opposed to events or artifacts. That presents both an advantage and a disadvantage. The advantage is that the researcher can make use of members of the population to build the sample. The disadvantage is that some members of these populations wish very much to remain unidentifiable and will, therefore, resist all efforts to find them. Such resistance makes representativeness difficult to achieve.

The most common and generally acceptable form of nonprobability sampling is referral sampling. **Referral sampling** begins with contacting some person or group in the population and using referrals to help identify and gain access to other members of the population. Two kinds of referral samples are **snowball samples**, which start with one or more *individual*

members of the population, and **network samples**, which start with an *organization* that contains members of the population.

With some populations the researcher or a colleague has special knowledge of elements in the population and can identify cases that would be particularly illustrative of the research topic, if not necessarily representative of the total population. Sampling these specially selected cases is called **purposive sampling**. In the case of studying families living off the grid, for example, the researcher might know of a handful of families doing so in significantly different ways or for significantly different reasons. Rather than using referrals, the researcher might concentrate on the select few. Another example of purposive sampling involves a two-stage strategy in which a large, probability sample who complete a survey. The survey might be designed, for example, to test a hypothesis about the relationship between personality characteristics and certain behaviors. The researchers then identify a subset of the respondents on the basis of their survey responses, then conduct in-depth interviews with this subset. You might, for example, choose individuals who had a combination of personality characteristics and behaviors that are theoretically interesting. The subset chosen for interviews is representative of unusual patterns of relationships, which may be of more interest to the research than the cases that fell in the middle ranges.

RELATIONSHIPS

As mentioned in the discussion above regarding time designs, this project will employ a longitudinal panel design. The panel study design is especially strong at capturing change measures, that is, measures of the difference between a value on a variable at one point and the value of the same variable taken earlier. So, for example, we might survey law students when they start law school and when they finish. At both times we include scales to assess their beliefs about the profession of law and their own sense of social responsibility or some other attitudinal measure. By comparing exit scores to entrance scores, we can see how students changed on those variables while they were in law school. It is reasonable to conclude that the change is related in some specific way to the experience of law school. The combination of panel design and change variable makes all the parts clear: a student had certain beliefs about the legal profession at Point A, after which she went through law school. By the end of law school, at Point B, her beliefs had changed. Establishing time order is one of the three necessary criteria for identifying causal relationships (see Introduction).

Given this information, however, we are reluctant to say that law school caused the change. We only know that the change coincided with attendance. We can use the survey to do two things to help us describe the causes of the change in beliefs more accurately. First, we can ask about other things that happened during that time, such as getting married, accumulating loan debt, having summer internships, death in the family, and so on. In the analysis stage, we can control for those other occurrences, called control variables. That

step gives substance to the criterion of **non-spuriousness**, which means ruling out plausible alternative explanations for the observed relationship. Second, we can ask many questions about the student's experiences in law school. What course of study, which clubs or activities, which instructors, perhaps experiences of discrimination or other mistreatment, and so on. These questions help identify the parts of legal education that are associated with more or less change on beliefs. By identifying all these **extraneous variables** or factors, we can more accurately claim that the change in beliefs during law school is more likely to occur under conditions X and Y than it is under conditions P and Q. That step puts context to the causal claims, specifying the conditions under which we observe that something causes something else.

In collecting data for Project One, you will ask certain questions about fellow students repeatedly. You may find increases in certain behaviors, such as calling home, in the midst of busier times academically. Or you may find decreases at those times. Because you can logically rule out the possibility that frequency of calling home causes the amount of academic work, you conclude that academic stress leads students to contact family. One particular advantage of the panel design is that you also know, on average, how often per week a student calls home. You are then in a good position to support the claim that academic stress causes an increase in family contact.

Even though this project is designed as a descriptive study, you still can hypothesize some findings. A **hypothesis** is a declarative statement about relationships among the concepts or variables in your study. It is a prediction about what your results will find. Writing hypotheses is a critical component of good methodological design because it gives direction to the kind of observations you need to make. The simplest form of a hypothesis contains an **independent variable** (the cause) and a **dependent variable** (the effect). The verb part of the sentence indicates the direction of the relationship (which is cause and which is effect) and the nature of the relationship (what kind of effect). A slightly more sophisticated hypothesis will also include **control variables**. Study these examples.

- At times of academic stress, undergraduate students call home more often than they do at times of regular academic workload.
    - Independent variable: Onset of academic stress
    - Dependent variable: Frequency of calling home
- Taking a student's major into account, at times of academic stress, undergraduate students call home more often than they do at times of regular academic workload.
    - Independent variable: Onset of academic stress
    - Dependent variable: Frequency of calling home
    - Control variable: Student's major

You will write hypotheses for this first project, then again for the other projects. Exercise Three is devoted exclusively to writing hypotheses and identifying their parts.

CONNECTIONS TO OTHER PROJECTS

As mentioned in the introduction, you will conduct two survey projects. Project Four will be a self-administered survey, so you will review many of the concepts related to designing survey questions then, along with the ethics of survey research. Furthermore, interviewing is a common component of field research. Depending on what you choose for Project Three, your field research may include some or no interviewing, and what is included will probably be more unstructured than what you will do for this project. Clearly, though, the conversational quality of the interaction in interviewing for the sake of a survey and for field research draws on the same observational behavior: asking good questions about what you are interested in and listening fully to the thoughts of others. In the field, the task of recording those responses poses more challenges than it does for face-to-face interviews, but the skills will translate from one kind of research strategy to another.

KEY TERMS

Anonymity
Attrition
Categorical measure
Change variable
Confidentiality
Control variable
Convenience sampling
Cross-sectional design or study
Data collection mode
Dependent variable
Dichotomous variable
Dummy variable
Element
Exhaustive
Extraneous variable
Face to face or in-person interview
Fixed sample
Hypothesis
Independent variable
Indicator
Informed consent
Interview
Interview schedule
Level of inquiry
Level of measurement
Longitudinal design
Mutually exclusive
Network sampling
Nominal measure
Non-coercion
Nonprobability sampling
Non-spuriousness
Ordinal measure
Panel study
Phone interview
Pilot testing
Placement
Population
Probability sampling
Protection of privacy
Purposive sampling
Questionnaire
Quota sampling
Ratio measure
Reactivity
Referral sampling
Repeated cross-sectional study
Researcher effect
Respondent
Sampling frame
Self-administered survey
Self-report
Semantic differential
Snowball sampling
Social Desirability
Study population
Survey
Target population
Time order design
Trend study
Variable

Project Four: Self-administered Survey

SKILL DEVELOPMENT FOCAL TASKS

- Designing a complete survey
- Pulling a representative probability sample
- Creating multiple indicator measures
- Selecting means for enhancing response rates
- Identifying appropriate control variables
- Securing informed consent
- Writing a cover letter

SURVEY OVERVIEW, PART 2

You were introduced in Project One to the general features of survey research:

- Surveys are essentially a question and response interaction which relies almost entirely on self-report as the indicator of all concepts.
- The participant in a survey is called a **respondent.**
- The two basic kinds of surveys are interviews and self-administered. These can be further distinguished in terms of the data collection modes (see below).
- Surveys offer many strengths as a research strategy:
    - They generalize well as long as they employ good sampling techniques.
    - Their reliance on self-report enables the researcher to operationalize opinions, attitudes, beliefs, and other interior states.
    - It is possible to operationalize dozens of concepts and test multiple hypotheses with a single survey administration. **Omnibus surveys,** such as the General Social Survey conducted by the National Opinion Research Center, cover a wide range of topics, not necessarily related to each other.
    - They are relatively cost-efficient, particularly self-administered surveys.
- Surveys carry a number of drawbacks and disadvantages:
    - They are inherently obtrusive, and therefore vulnerable to reactivity.
    - The connection between self-reported behavior and actual behavior is often weak (due to the influence of social desirability), limiting the usefulness of surveys in predicting behavior.
    - Their reliance on verbal interaction renders them less useful with any population for whom language is difficult.
- Measurement validity presents challenges for survey research, which can be met by carefully pilot-tested instrument design and the use of multi-item scales.

- Establishing causality with survey research presents challenges as well, often requiring longitudinal designs.
- In addition to the absolute necessity of good sampling techniques, survey research consists of both an instrument and an administration strategy: asking the right questions of the right people in the right way.

Data Collection Modes

In order to understand survey administration, it is important to examine both the characteristics of high quality surveys and the nature of the many judgment calls involved. Before examining the features of surveys in greater detail, let us briefly review the different data collection modes. As you will see below, data collection mode affects the quality of those features significantly.

The three kinds of **interview data collection modes** include:

- In-person face to face
- Video face to face
- Phone

The three kinds of **self-administered data collection modes** include:

- Mail-in
- Hand-in or group administered
- Web-based

The advantages and disadvantages of interviews were reviewed in Project One. Here we review the different types of self-administered surveys. Most of us are familiar with **mail-in** surveys. They are mailed out to the sample and returned by mail. **Hand-in** includes two variations, both of which involve the researcher delivering the instrument in person and remaining available for respondents to return their completed questionnaires. One option is administering the survey to a group of people who are gathered someplace, either for the express purpose of completing the survey or for some other reason which is pre-arranged to include the survey. This **group-administered** technique is useful with populations like students, work or sports teams, organization members, or employees who already gather on a regular basis in set places. The other type of hand-in survey uses a locating rule of inclusion for the sampling frame, such as persons attending a particular sporting event or using some services at a particular time. It could include people who are waiting for something else to start, for example, a ball game, a parade, or township offices to open. (It is important to note that a sampling frame like that tends to represent a population too loosely to be considered a probability method, thereby losing one of the major advantages of survey research.)

Finally, **on-line** or **web-based** surveys typically rely on e-mail and Internet browsers, with four major variations, and a fifth hybrid option:

- The survey is in the body of the email itself, and the respondent replies to the email after answering the questions without having to go anywhere else on the Internet. This option is distinctly an **email survey**, even though many people use that term to mean generically web-based surveys which use email invitations.
- The invitation to participate is sent out by e-mail with a link (URL) to the survey itself.
- The survey is posted on the web and anyone who happens to visit the host web site is welcome to complete the survey.
- Both invitation and survey are contained in the e-mail or in an electronic attachment, which is then returned electronically.
- A fifth option is to administer the survey in both paper and web-based forms by sending the survey in regular mail and including the URL in a cover letter.

Giving your respondents the flexibility of choosing the mode they prefer generally improves response rates. Software developments are making web-based data collection more popular, but it should be kept in mind that this mode is appropriate only for limited populations.

As you will see shortly, the self-administered approach has the advantage of reducing reactivity to the person of the researcher (there might still be reactivity to the instrument). It also tends to cost less than interviews because person-hours are not needed to record the responses. Even with group-administered, the labor investment of having one person coordinate distribution and collection of surveys to a group of people is relatively low. The primary disadvantage is that response rates tend to be lower for self-administered surveys than for interviews.

With this overview complete, let us turn the presentation to more details on features of administration. An elaboration on instrument construction will follow.

SURVEY ADMINISTRATION FEATURES

Per unit cost. How much cost is involved in producing a single case, that is, a completed survey? Depending on the data collection mode, these costs could include instrument design, personnel, supervision, printing and supplies, data entry, mailing, phone calls, travel, and so on, as well as the costs of unsuccessful attempts to contact possible respondents. In calculating per unit costs, you need to distinguish between start-up or one-time costs (like instrument design and training) and ongoing costs (like data entry and phone calls). For example, if you expect the survey to be administered repeatedly, it might be worth it to invest more in the design of the instrument (a start-up cost) so that you can save money on data entry (an ongoing cost). By contrast, if only one administration is planned, it is not as efficient to invest a lot of money up front in the hope of reducing ongoing costs.

Time. How long does it take, once the instrument is ready, to collect the data? This includes the total out-and-back time plus data entry for the entire sample, or until reaching

the point of diminishing returns (when continued effort produces barely any more completed surveys).

Sampling frame construction. The construction of the sampling frame obviously depends on the type of population, but also on the data collection mode. Face to face interviews require addresses or a locating rule of inclusion (people in a certain place at a certain time), phone interviews require phone numbers, mail-in requires addresses, hand-in or group administered requires a rule of inclusion, and web-based requires e-mail addresses and/or mailing addresses, depending on how you want to make the initial contact.

Vulnerability to sampling frame error. Recall that the difference between the target population and the study population is usually due to defects in the sampling frame. Are the mailing addresses, e-mail addresses, or phone numbers correct? Is the list complete?

Sample quality. Given the type of sampling frame needed to represent various populations, how good a sample can we expect to draw? This refers only to the quality of the sample that can be drawn, not to the final collection of cases, which depends on the number of people who complete a survey (**response rate**).

Overall response rate. Of the cases selected for inclusion in the study (the sample), what proportion produces completed surveys? This is a characteristic on which the data collection modes differ markedly and could be a determining factor in your choice. Response rates can be calculated based on the number of respondents who complete the whole survey (**full response rate**) or the number who complete at least some of it, preferably the parts most important to the study (**partial response rate**). Two factors are closely related to response rate. The first is *the ease of follow-up procedures*. That is, if a person does not respond or is not available at the first contact attempt, how easy is it to try again? Generally, three attempts is the limit—more than three seldom increases the chance of a response substantially, but may be worth it, particularly if you are close to your target response rate. What kind of time and other resources does it take to make multiple attempts? The second related factor is *the ease of determining* ***non-response bias***. That is, how accurately and with how much detail can we describe non-respondents to see if they differ systematically from the respondents? The more information we have in the sampling frame, the more accurately we can estimate non-response bias.

Item completion rate. Of the returned surveys, what proportion of the items on the questionnaire was completed? Although length (total number of questions asked) is a factor in item completion rate, a more important factor is how the researcher presents the survey to respondents. The key components are soliciting cooperation and assurances of ethical and discrete uses of the information provided.

Nonreactivity. Ultimately, we want to achieve nonreactivity, but, given the self-report nature of surveys, a more realistic goal is simply to reduce reactivity. A respondent might react to the person of the researcher or to the asking of the questions. Both would undermine the validity of the measures.

Quality of the questions. One characteristic is flexibility. Can the questions be adapted to the needs of the respondent or to unexpected revelations from the respondent? How much does changing questions during administration compromise comparability across cases? Another characteristic is nuance and depth. How complicated can the questions get? To what depth can they delve? How well can the questions cover sensitive topics that may require some sense of trust between researcher and respondent?

TABLE 4.1. COMPARING DATA COLLECTION MODES

	Interviews			Self-administered		
Features	Face to Face	Phone	Video	Mail-in	Hand-in	Web-based
Per unit cost	Very high	Relatively high	Relatively high	Relatively low	Low, but varies	Varies greatly
Time (out and back)	High, but varies and under control	Low to medium	Low to medium	High, not in control	Very low	Low, but varies
Sampling frame (SF) construction	Relatively easy, but varies; also flexible	Very easy, for simple random; Hard for stratified	Difficult, but varies greatly by population	Difficult, but makes stratifying possible	Easy	Difficult, but varies greatly by population
Vulnerability to SF errors	Low	Low	Varies greatly by population	Somewhat high	Very high	Varies greatly by population
Sample quality (representative-ness)	High, but varies somewhat by population	Mixture of strengths and weaknesses)	Low, except for very particular populations	Relatively high	Varies by population	Varies greatly by population
Overall response rate	Very high	Medium to high	Medium to high	Low	Medium to high	Low to medium
Follow-up procedures	Controllable, but demanding	Controllable and easy	Controllable and easy	Controllable and easy	Difficult to impossible	Controllable and easy
Determining non-response bias	Very easy	Very difficult	Difficult, but depends on SF	Easy, but depends on SF	Very easy	Difficult, but depends on SF
Item completion rate	Very high	High	High	Medium to low	Medium	Medium to high
Non-reactivity						
Researcher related	Highly problematic	Somewhat problematic	Highly problematic	Low	Low, but can be problematic	Low
Instrument related	Relatively low	Relatively low	Relatively low	High, but can be minimized	High, but can be minimized	High, but can be minimized
Quality of Qs						
Flexibility	Very high	High	Very high	Very low	Very low	Low to medium
Depth and complexity	Very high	Very low	Medium to high	Low to medium	Low to medium	Low to medium
Sensitivity (personalness of questions)	Problematic, but possibly easy	Problematic	Problematic	Can be accomplished, with limits	Can be accomplished, with limits	Can be accomplished, with limits
Quality control						
Ease of supervision	Costly, but possible	Very easy	Very easy	Not applicable	Not applicable	Not applicable
Ensuring uniformity	Costly, but possible	Less costly, fairly easy	Less costly, fairly easy	Very easy	Very easy	Very easy
Ensuring quality	Costly, but achievable	Less costly, achievable	Costly, but achievable	No control	Difficult, but attemptable	No control
Assuring respondent of anonymity	Not possible	Not possible, practically speaking	Not possible	Very possible	Possible, some doubts may arise	Very possible, some doubts may arise

Quality control. For the most part, quality control (consistency across cases) is an issue with interviews but not with self-administered surveys, because self-administered surveys require little interaction with respondents. In order to ensure quality with interviewing, supervision of the interviewers is often required. That raises the question of how easily the interview can be supervised. With self-administered surveys, quality control is a function of how well the instrument is designed.

Anonymity. As you know, confidentiality should always be insured; anonymity, wherein the respondent's identity is not known even to the researcher, is a design decision, sometimes invoked in order to encourage respondents to answer questions with complete honesty. Recall that anonymity is not simply whether the respondent is identifiable, but whether the respondent believes his or her identity is and will remain completely masked.

Table 4.1 summarizes the ways that the data collection mode interacts with the numerous characteristics of survey administration.

MEASUREMENT ISSUES AND INSTRUMENT CONSTRUCTION

Because survey research relies so heavily on self-report to operationalize all concepts, measurement validity depends wholly on instrument construction. Furthermore, the wording of questions and responses, the presentation of instructions, and the sequencing of items contribute substantially to reliability. In order to design good **questionnaires** (the term for self-administered survey instruments), therefore, the researcher must be able to make use of a wide range of tools and options with creativity, discipline, judgment, and a keen sense of the target audience. This section will review the basic elements: types of questions, types of responses, placement and instructions. The following section will address three additional elements: scales, other report and aggregated self-report, and enhancing response rates.

Questions

Question wording. *The primary and unwavering criterion for a good survey question is clarity*, which means that it elicits the information you want to collect. Most of the time that requires wording that is exact, yet simple—but the bottom line is that respondents understand what you are asking for. Sometimes I have found that making a question very precise actually makes it harder to understand, so be careful to pilot test everything. In fact, the best way to pilot test surveys is to administer them to a group and to follow questionnaire completion immediately with a group discussion. During the discussion a member of the research team asks respondents how they understood the questions, whether anything was unclear or confusing, and how they felt about answering the questions. That feedback informs the next set of revisions, which are again pilot tested.

One of the trickiest parts of instrument design is achieving just the right level of breadth and specificity. Consider the following two items:

- Overall, how satisfied were you with hospital services during your stay?
- How often do you get in trouble with your parents?

Both questions are worded broadly. Perhaps you want respondents to bundle together a variety of experiences that constitute "hospital services" or "getting in trouble." Perhaps not, though, and you need to make sure that the answers you get will allow you to operationalize the corresponding concepts.

Broadness in wording is sometimes appropriate, but ambiguity and vagueness will hardly ever serve you well. In many work situations, for example, "management" can refer to several layers of an organization, so if you want respondents to share their thoughts about management, specify what that means. As another example, in my research on college environments and outcomes, I have discovered that the term "studying" means many different things to students. In studies of the ways that couples share household duties, terms like "work around the house" and "childcare" would have to be defined. These are just a few examples. Your literature review should familiarize you with the concepts that are prone to ambiguity, as well as provide some ideas for you to sharpen the meaning of the words you use.

Beyond clarity of meaning, two other wording requirements should be met. The first is **singularity**. Each item should have one and only one idea behind it. You can detect this best when it's not present, as in **double barreled wording**, which is common with attitudinal measures. For example, some respondents would have difficulty deciding whether they agree or disagree with the statement, "Gun control legislation and gun purchase background checks would help reduce crime" for two reasons. First, they might feel differently about gun control and background checks, and, second, they might see different connections between those two approaches and crime reduction. Another common type of double-barreled question, also in the form of agree-disagree, is when a behavior is presented along with a reason. For example, "I enjoy meeting people with backgrounds different from mine because I learn a lot about myself by interacting with them." A person may enjoy meeting diverse people, but for some other reason.

Along with clarity and singularity, good item writing demands **neutrality**. Items should be worded in such a way that respondents feel free to answer according to their own positions rather than feeling that the item suggests a "correct" answer. Non-neutrality is often a problem with political polling that is sponsored by a partisan group. You may well have encountered both of the main variations on non-neutrality. A **leading question** suggests to the respondent, usually subtly, that a certain response is more acceptable than the alternatives. Consider the following examples of leading items in agree-disagree statements. Notice that, since these are agree-disagree statements, the statements themselves state a position. That is not the problem. By definition, an agree-disagree

statement must be worded in one direction or the other. *The problem is when the wording implies that either agreement or disagreement is the more reasonable response.*

> The killing of human babies by abortionists should be outlawed and violators prosecuted as the murderers they are.
>
> Alternative: Abortion should be made illegal.
>
> Since we have an enormous deficit already, tax cuts for the rich are the stupidest idea in Washington.
>
> Less bad, but still a problem: Even though we have budget deficits, taxes should be cut.
>
> Alternative: Tax cuts are appropriate at this time.
>
> The tenure system saddles college students with professors who are useless in the classroom.
>
> Alternative: The tenure system undermines the quality of college instruction.

Loaded wording, the second variation on non-neutrality, puts too much emotional charge in the item, usually intentionally and usually but not always in order to influence the response. The examples above have a number of loaded words: abortionists, murderers, enormous, stupidest, saddles, useless. Now, you may be thinking, for example, "Some professors are useless!" And you likely have your own opinions about abortion, tax cuts, and many other issues. *As a researcher*, however, you need to keep those judgments out of instrument construction. Instead of words like "stupid" or "ridiculous," use words like "ineffective" or word items so that respondents can agree or disagree with a position. Instrument construction is not an exercise in creative writing or persuasion; it is a data collection exercise.

Compared to the rather conspicuous violations in self-administered surveys, problems with neutrality and clarity *in interviews* can be harder to detect and more dangerous to validity. That is because the instrument is the combination of interview questions and the interviewer's presentation of self, so reactivity enters the interaction. Note in the chart above which compares data collection modes that the quality control factors are all labeled as costly in face-to-face interviews. Good training of interviewers should reduce neutrality and clarity problems, but the only way to make sure is to have another member of the research team observe the interview. Ideally this would be done unobtrusively, but that is usually impractical (it can be done with phone interviews fairly easily by having a supervisor listen in on the phone call).

Besides maintaining clarity and neutrality, questions serve the purpose of keeping the respondent engaged in the survey. To do so, the questions should be relevant to the target population and varied in form. To get a sense of variety in questions, it helps to think about response categories. Let us turn our attention to that element of instrument construction.

Question functions. Note that questions on a questionnaire serve several purposes. The primary purpose is to solicit a response that contributes to the construction of some

measure needed to answer the research question. In addition, you will often need some questions in order to direct the respondent's movement through the questionnaire. The rationale for directing respondents through a questionnaire in tailored progression (not everyone answers every question) is called the **skip logic**. Careful use of skip logic is extremely valuable because it allows you to take advantage of the flexibility of instrument design to address a wide range of research questions while keeping the survey as short as possible for each respondent. If certain questions should be answered only by some of the respondents, you need a **filter question.** For example, if you want to know what college students are doing after they graduate, you might ask a general filter question about whether they plan to enter the workforce, continue their education, do community service or serve in the military. For web-based surveys, it is also possible to base skip logic on information that you load into the survey software. You might, for example, know whether an employee is salaried or hourly, with different items on an employee satisfaction survey tailored to each type of worker. You do not have to ask them which category they belong to, but show only the relevant questions to each respondent. Depending on how they answer the filter question or the characteristic they have on a filtering data field, another set of questions, called **contingent questions,** would address the specifics relevant to their personal plans. You can easily think of other useful filter questions, such as whether a respondent has children and, if so, how many and what ages, or whether a respondent has supervisory responsibilities or not.

A slightly different use of a filter question allows a respondent to select first from a set of experiences or preferences, which then become the specific items in a following list which the respondent is asked to evaluate or otherwise rate. This is a feature of some web-based survey software, but could not be accomplished on paper. For example, you could ask an outgoing hospital patient which of a lengthy list of hospital services he or she used during the stay. On the next page of the survey, only the selected services would appear in a list that the patient is asked to rate for quality. This is called **piping**. If the list of possible services or experiences is less than ten or so, piping might not be worth it (because it involves answering one question just to determine how to answer another question), but if you do not pipe, you have to make sure that "Not Applicable" is in your response options.

Sometimes, particularly in exploratory stages of some survey studies, you want to know how respondents understand your questions. One way to handle this is through the **cognitive interview,** which is a group-administered technique where the researcher is present while the questionnaire is being completed, after which he or she asks the group how they interpreted the key questions. When that is not feasible or when you simply want to learn about the respondent's understanding of the question to add context to your analysis, you can put **interpretive questions** on your questionnaire. This would usually be in an open-ended format. After asking about amount of quality time someone spends with their teenage children, you could ask something like, "What is your idea of quality time?"

Responses

In the discussion of instrument construction for Project One, the merits of open- and closed-ended questions were described. You may find it helpful to review that discussion. For now, recall that **closed-ended, or multiple choice**, items have a pre-determined set of response options, whereas **open-ended** items allow the respondent to answer in free text. Closed-ended items are easier to answer and to analyze, but more limiting to the respondent. Open-ended items presume less about what respondents might have to say. but make more demands of both the respondent and the researcher.

As stated earlier, closed-ended questions present a wide range of possibilities. Before reviewing some examples of closed-ended types, let me emphasize three features of these response sets that apply to all types.

First, any set of closed-ended responses should be **exhaustive and mutually exclusive**. That means that all possible categories are listed and that a respondent will fit into one and only one category (see Project One). For religious affiliation, do not list both Lutheran and Protestant, as those categories overlap. Because a truly exhaustive list for some categories would be impractically long (occupations, for example), it is best to list the categories likely to fit most of your cases, then include the category "Other." You may or may not want a fill-in space by that category (e.g., "Other:_____"). It is also necessary sometimes to include the option, "None." With ordinal scales (especially amounts of things) you do not need to worry about making your response options exhaustive because you can always put the top range as "20 or more" or "$150,000+."

A second issue in crafting response categories is the use of a **Neutral or No Opinion option**. A list of response categories without a middle position is considered a **forced-choice question**. On agree-disagree and satisfaction-dissatisfaction rating items the neutral category is explicitly identified. With other response categories, you can create a neutral option by having an odd number of points on the scale or create a forced choice by having an even number of points (there is no middle on a 1-6 scale, for example, but there is a middle on a 1-5 scale). As the term suggests, absence of a middle position forces respondents to place themselves on one side of an issue or another, and some people are honestly neutral on some issues. Would you rather that those people skip the item or slightly misrepresent their true opinion? At the same time, the neutral option allows some respondents to choose an easy, noncommittal answer when they probably do have an opinion if they give the matter enough thought. The use of neutral-inclusive and forced choice response categories should be considered carefully with respect to both the nature of the target population and the uses to which your survey data will be put.

Third, it is increasingly important to consider the option of including the response category, **Prefer not to answer**. Listing that option, of course, invites respondents to think about whether they prefer not to answer when maybe they wouldn't have thought about the sensitivity of the item if you did not include that option. So it is risky. There are, however, two compelling reasons for keeping this possibility in your toolkit. Even when your

instructions indicate that respondents can skip items, you can comply with the ethical requirement of non-coercion more explicitly if you include this response option on particularly sensitive items. The other reason is that, with online surveys, you sometimes have to make some questions required for skip logic to work; "Prefer not to answer" satisfies the software requirement without actually forcing someone to provide an answer.

Having reviewed those universal issues for response options, let us look at some types. The following list is suggestive, not comprehensive.

Likert scales. Agree-Disagree formats which present a statement so that the respondent can indicate how he or she feels about it (including "strongly" one way or the other). This response option requires careful consideration of the neutral/forced choice issue.

Amounts, durations, and frequencies. Sometimes we need to know how much, how often, how long, and so on. How many times has something happened? How often does something occur? How much time per week is spent on certain activities? Response categories might be in ranges (e.g., less than five, 6-10, 11-15, 16-20, 21 or more) or there might be room to fill in exact numbers (e.g., To how many law schools did you apply?), although that is uncommon because it presents recall accuracy problems. Frequencies can also be measured using generic terms like frequently, occasionally, seldom, or never. Keep in mind how you will eventually operationalize your measure and whether ordinal or ratio level of measurement will suffice for your analysis.

When asking about quantities, it is important to provide the option for none or zero or not at all, if you want to be able to distinguish between small amounts and nothing. Consider two types of events about which you are asking frequency of occurrence in a college student's life: (a) How much time in a typical week do you spend watching television? (b) How often have you been a defendant in disciplinary hearings? With regard to watching TV, it probably would not matter much in analysis whether someone watched absolutely no TV or watched thirty minutes per week. On the other hand, it makes a big difference whether someone has had no disciplinary hearings or one. Know your target population and anticipate measurement validity issues.

Ratings. You can ask for people's ratings of all kinds of things. You might want them to rate the quality of goods or services, their satisfaction with something, or the importance of various goals or objectives. Examples of ratings-related response categories include poor, fair, good, excellent; not important, somewhat important, very important, essential; and satisfied to dissatisfied.

Semantic Differential. When you are asking respondents to give you their impression of something or someone, semantic differential scales might help. These scales set two opposite words on either end of a continuum and ask the respondent to show where their impression falls on the continuum. For example, you might name a public figure and have respondents indicate where on a scale of Decisive to Indecisive they would place that figure. Or the scale could be Effective-Ineffective, Charismatic-Dull, or Trustworthy-Untrustworthy.

You could ask people to rate the department of their major on scales of Warm-Cold, Friendly-Unfriendly, Helpful-Unhelpful. Use your imagination.

Descriptive categories. You might want to know what category a person belongs to, such as ethnicity, religious affiliation, political party, highest degree attained, major, industry in which they work, job title, marital status, or personality type, just to mention a few. For this kind of question, it is important to keep in mind the rule that response options be exhaustive and mutually exclusive.

Placement

The basics of item placement were discussed when you were planning the first project, the face to face interview survey. To repeat, simple yet engaging items go at the beginning, items that do not require serious recall go at the end, and the middle should follow one of a few types of logic.

- If there is some time sequence to the experiences referred to in your questions, use a chronological order for items in the middle. This category includes sequences such as early college through later college years, pregnancy through labor and delivery through early infancy, check-in to a hospital through treatment through discharge, and so on.
- If your questions involve sensitive topics, the middle should progress from less sensitive to more sensitive items. This category could include questions about deviant behaviors, traumas such a sexual assault or domestic violence, the effects of debt or unemployment, and so on.
- If your questions center on evaluation of experiences or services, you can use one of two approaches. A **funnel sequence** moves from the general to the more specific, and allows the respondent to think about the big picture first, and, as more specific features of the experience are raised, have some overall context in which to consider the details. For example, you could ask graduating students how satisfied they are with their overall college experience, then follow with satisfaction ratings of specific aspects of that experience. An **inverted funnel sequence** moves from the specific to the general. In this case, you are allowing the respondent to think about particular aspects of the experience, gradually building to the big picture. For example, student feedback on teaching often has items to rate the instructor's availability, respect for students, organization, etc., then an item for overall effectiveness.

As the description of funnel sequences suggests, placement is not only about beginning, middle and end, but also about how items are placed in relation to each other. This has a few implications. First, you are not controlling respondents' minds, but you are providing context for each item with the items that come before it. Consider this example of a list of life goals, the importance of which you are asked to rate for you personally:

- Raising a family
- Achieving a sufficient standard of living
- Being respected for your accomplishments
- Protecting the environment
- Doing works of charity

If you were to move the standard of living item down the list so that it came after the environment and charity items, it would have a different meaning in many individuals' mind from its meaning in relation to raising a family. The same would apply to the accomplishments item, and many other permutations as well. Of course, you have to put all of your survey items somewhere and, as mentioned, you do not want to be in the business of manipulating responses. One option is to randomize the order. Most web-based survey software allows you to do this during the administration of the survey (different respondents get the list in different orders), but if it is a print survey, by random I mean shuffle the items and let them each fall where they may. You could also try to achieve a certain logical order. In a list like this, that logic might go from day to day concerns at the beginning of the list to broader, long-term concerns at the end, or vice versa. There are no clear-cut rules for this, although I see great advantages to randomization if you're using web-based delivery.

Whatever you do choose, the second important implication of this is that you should not alter the order of items if you are administering the survey in multiple phases longitudinally. The results are difficult to compare across time points if the order has changed. If the list of life goals above is administered to a panel sample in a set order, for example, when you conduct a follow-up study five years later, they should be in the exact same order. If you decide to add items in the second study, they should go at the end of the list.

Beyond single item placement, whole batteries of questionnaire items also are subject to sequencing effects. Are respondents asked about educational loans before or after an overall college satisfaction item? Are respondents asked about international political issues before or after domestic issues? Whatever the topic of your survey, reflect carefully on the effects of order in item placement.

Instructions and Layout

It is easy to overlook the importance of writing instructions when you are working extremely hard on item wording. I have a simple twofold rule for writing instructions for questionnaires: Be absolutely clear and concise in your instructions and then design the survey as if no one will read them. They do indeed have to be precise, although you want to avoid stiff formality in the grammar. For example, if a phrase sounds better ending in a preposition, let it be. You probably should avoid slang, but informal wording is fine. You have to know your target population well.

General instructions should be provided at the very beginning of the questionnaire, then place section instructions along the way as needed. If you have **skip patterns**, where

respondents might not have to answer some questions depending on responses to previous items, use layout to help the respondent navigate the questionnaire, with things like arrows, indentations, and shading. If the survey is lengthy, think about putting something at the top of the last page or two which says something like, "Now, just a few final questions about yourself."

SPECIAL TOPICS IN INSTRUMENT CONSTRUCTION AND ADMINISTRATION

Scales

As stated above, surveys afford the researcher the opportunity to take advantage of self-report. Attitudes, personality traits, dispositions and many other complex concepts can therefore be operationalized effectively with good survey design. Precisely because they are complex concepts, careful thought must go into selecting the best indicators. Typically, these concepts are operationalized through scales or indices. A **scale or index** is a set of items, the responses to which are combined to create a value for a respondent on a given variable. When scales are used to capture the many components of multi-dimensional social reality, the terms **concept** and **construct** are used interchangeably. There are a number of reasons that scales work effectively for this purpose. Before examining those reasons, however, let us first consider the application of measurement validity to scales.

In earlier projects, you had to achieve face validity, which meant that you presumed the thing you could observe was a direct manifestation of the variable you were trying to capture. If you ask people how many children they have, they may or may not tell the truth, but you cannot do much other than use the number they give you. Of course, you do have to specify clearly what you mean by "how many children do you have?" because of factors like adoption, step families, and grown children not at home. But ultimately you have to take their answer at face value. Now what if you wanted to capture a concept like whether they like being a parent. Can you ask, "Do you like having children?" and accept the response at face value? Maybe, but having children is a many-faceted condition and a person's answer might depend on which part of parenting you mean. That's where scales come in handy, which will be explained further shortly. The point is that you need a more robust type of validity to make sure you have captured "satisfaction with parenting" accurately.

Construct validity assumes that groups of concepts are theoretically related to each other. Familiarity with relevant theory will come from your literature review. The development of valid scales requires simultaneous measurement of the related concepts. In fact, scale development is usually a research project in itself so that when you do want to use a scale in another research project (usually surveys or experiments), it is already validated and, therefore, it is not necessary to measure all of the theoretically related concepts as well, which could make the survey or assessment unnecessarily long.

Furthermore, theory will often suggest that the concept should be higher in certain populations and lower in others. Scale development is also tested on many populations.

In short, the concept should be correlated positively but not too strongly with similar concepts, negatively but not too strongly with dissimilar concepts, and no correlation with totally unrelated concepts. Why "not too strongly"? If your measurement of a concept is correlated too highly with another concept, you have not captured its distinctive features. Imagine, for example, that your measure of parenting satisfaction is very highly related with self-esteem. Maybe you have not captured anything more about a person's feelings about parenting other than whether they generally like themselves and, therefore, are satisfied with most of what they do. To draw on another example, you would expect a measure of social responsibility to correlate negatively with narcissism or rugged individualism. Regarding the "no correlation" part of this, theory again will provide guidance regarding what concepts should be independent of your measure. Parenting satisfaction should not be correlated with SAT scores or weight, for example.

The second condition is that your measure should vary in predictable ways across populations. You would want your parenting satisfaction measure to be higher among those who choose to have large families than it is among those who choose to have small families. A social responsibility scale should show higher scores among Peace Corps volunteers than it shows for violent criminals. At the same time, you would expect that parenting satisfaction does not differ across people of different religions or the sex of the children, among other things.

Now let's suppose that you want to measure the concept of "quality of parenting." This is not an inner disposition, but an ability. Once again, a literature review will provide you with many ideas about what makes for good parenting. As will be explained below, you would probably create a combinative scale for this concept, one that refers to a number of aspects of parenting such as time spent with the children, care for their physical and emotional needs, attention to schoolwork, reliability in keeping schedules and appointments, etc. Because you are trying to capture the extent to which a person has a capacity to do something, the kind of validity you want to achieve is **criterion validity**. If that capacity is something the person has not yet demonstrated and you want to see whether he or she is likely to be good at it, your measure should have **predictive concurrent validity**. You might administer an assessment in parenting or marriage preparation classes, then validate that measure against how well the students in the class actually parent when they do have children. You could validate it by comparing the parenting quality measure with an assessment by a case worker based on home visits. (To use examples that you can easily relate to, the SAT and ACT are supposed to have high predictive criterion validity for success in college, and a driving test should have high predictive criterion validity for one's ability to operate a motor vehicle.) If the ability is something the person has at the same time as the measure is taken, then the measure has **concurrent criterion validity**. Testing for that type of validity would require verifying the self-reported measure of the capacity against another

assessment of the same thing. Back to the social responsibility example, if a person scores high on that scale, we would expect that he or she also exhibits—perhaps by self-report as well—a number of behaviors that demonstrate a commitment to improving the world around.

Now we return to the idea that scales work well to measure complex concepts. Why? First, some concepts, particularly those with high social desirability (or undesirability), are best observed indirectly. Imagine asking someone, for example, "How prejudiced are you?" or "How religious are you?" You would get answers, but how much would you want to rely on the answer? Instead, it works better to consider the many manifestations of prejudice or religiosity and ask people about those things. Moreover, there are some things that people might not be fully aware of about themselves, such as personality type or gender identity. Again, identifying traits or behaviors that, taken together, constitute those underlying characteristics is preferred to a single, direct question.

Second, multi-item indicators eliminate some of the error introduced by the slightly different ways that people interpret words and phrases, a phenomenon known as **idiosyncratic variation**. To use the example of religiosity, you would probably want to include items that refer to worship, religious artifacts, and relationship with God or the Almighty. But individuals have different ways of incorporating those elements into their religious experience. If you limited your measure to one or two of those elements worded in specific ways that resonated with some and not with others, you would likely miss the mark for some unknown number of respondents. What scaling does is to broaden the possible inputs so that, overall, you are more likely to touch on a higher proportion of the construct's elements in ways that respondents understand as intended.

Third, scales allow the researcher to check the reliability of a measure by examining the correlations between all of the pairs of items in the scale and the correlation of each item with the combination of all of the others. These **inter-item** and **item-scale correlations** provide very robust reliability checks.

The type of scales implied by the discussion above is called **iterative**, that is, all of the items are essentially repeated measures of more or less the same thing. Another type of scale is also used to enhance validity and reliability, called a **combinative scale**. A combinative scale consists of items which are all different expressions of the underlying concept. For example, if you were trying to measure overall study time, you might ask about hours per week spent in a typical week doing a number of things such as lab exercises, reading, preparing for exams, working on group presentations, etc. You would not expect the responses to these items to be correlated with each other, and it might make perfect sense for some people to spend no time on one of the activities and lots of time on others. If you saw those patterns on an iterative scale, you would be concerned, but it is not a problem for a combinative scale.

The use of scales underscores a step often necessary for full operationalization: **rules of combination**. Whenever you operationalize a single concept with multi-item indicators, you

must specify the exact procedures for combining the values from each item into the single value for the concept. Consider a simple example. You have ten agree-disagree statements as indicators for a construct called Relationship Maintenance Needs, which is defined as the amount of care and attention a person needs in a dating/romantic relationship. For all ten items, agreement with the statement indicates high maintenance. Your rule of combination involves two steps. (1) The ten responses are numerically coded (0 for strongly disagree through 4 for strongly agree), then (2) those ten values are averaged to get each respondent's Relationship Maintenance Needs score. Now let's consider some variations on the base form.

- If agreement with some of the statements indicates low maintenance—that is, they are negatively worded—you must first code the positively worded items as above (4=strongly agree, 0=strongly disagree), then reverse code the negatively worded items (4=strongly disagree, 0=strongly agree), then average all of those properly-coded values. This is called **reverse coding**.
- Suppose five of the items for Relationship Maintenance Needs come from a set of agree-disagree statements, and five come from a self-rating scale. The former uses a five-point scale and the latter uses a four-point scale, like so: Not at all like me (0), Somewhat like me (1), Quite a bit like me (2), and Exactly like me (3). All of the items are worded so that agreement or a positive rating correspond with high maintenance. It may seem that the solution is again to average all of the items, but if you did that, you would place greater emphasis on the agree-disagree items than on the self-ratings because 5-point scales have more variance than 4-point scales. In addition, if, by chance one person strongly agreed with all of the statements, but marked Not like me on all of the self-ratings, that person would have an average of 2.0. If someone else did the reverse, that person's average would be 1.6. Both have half of the characteristics of High Maintenance, but one scored higher than the other. So you should recode the self-ratings to 0, 1.2, 2.4, 3.6, and 4.0. this is called **rescaling** or **standardizing response ranges**.
- Suppose you have some items among your indicators that are strongly associated with your concept of High Maintenance and others that are positively associated, but less strongly. In other words, some of the indicators carry more weight than others. You might want to consider assigning different values to the agree-disagree response options for those items. Perhaps those stronger items are coded 0, 2, 4, 6, and 8, instead of 0-4. To step aside momentarily from this example, suppose you were measuring relational violence. You wouldn't want to have a "slap" carry the same weight as an "attack with a deadly weapon" in your overall construct. Making these adjustments is called **weighting the indicators**.

In the example above, the final combination of item values was averaging. It is also common to take the sum of the item values. If the response options for index items include Yes and No, your rule of combination might be to take the percentage Yes responses. Those

are the three most common arithmetic operations used in computing scale or index scores. Other operations are sometimes used for other rules of combination, which will be discussed below.

The Special Uses of Other-report and Aggregated Self-report

At the beginning of this chapter it was stated that surveys rely "almost entirely" on self-reported indicators. Why not "entirely" when, after all, a survey is someone answering questions? The reason is that some surveys are designed to use **other-report** to operationalize concepts. Other-report resembles self-report in that a person provides the information that constitutes the observation, but, as the label suggest, the information is provided about someone or something other than the respondent. Two uses of other-report are (1) when a person is familiar with another person (**familiarity other-report**) and (2) when a person represents a group or organization (**representative other-report**). Familiarity other-report can take the form of a parent describing a child, a teacher describing a student, or a manager describing the workers in her unit, just to name a few possibilities. The person completing the questionnaire or interview might be in a position to provide a great deal of information about the "other," but it is important not to assume more familiarity than is warranted.

Securing informed consent is challenging with familiarity other-report. When used to collect information about minors from their parents, the parents' submission of the survey is construed as informed consent. If you are collecting information about students from a teacher, you must obtain the consent of the students or, if they are minors, from their parents. Most likely, this kind of data collection would be part of a larger, multi-pronged approach for which you would need the consent of the students and their parents regardless of the use of other-report. In the case of one adult providing information about other adults with whom she works and perhaps over whom she has supervisory authority, the procedure for obtaining informed consent depends on the kind of information collected. If individuals are not identified and the questions do not present privacy violations, it is probably not necessary to obtain consent from the people about whom information is being collected. Otherwise, it is best to obtain informed consent and to allow individuals the option not to participate.

Representative other-report is very common when your research unit of analysis is a group or organization. In that case, you would administer the survey to a representative of the organization asking him or her to answer questions about the organization. Businesses all over the country submit monthly reports on employment figures and economic activity. Eventually those data become part of government archives and are accessed later as artifacts (government records), but the initial step of data collection is a survey utilizing other report. In addition to businesses, schools, churches and other places of worship, volunteer agencies, and so on can be contacted to gather information. Instrument construction and survey administration for other-report designs are, for the most part,

identical to the practices used for self-report designs. The instructions should carefully define the information needs of the project. Usually the instrument includes a section for the person completing the questionnaire to provide information about himself or herself, but it is limited. Rules of combination should be specified as precisely for other-report as they are for self-report.

One other useful and challenging variation on self-report in survey research is **aggregated self-report**. This technique is used when you want to use input from members of a group to create a measure of an individual or organization affiliated with that group. Although this may sound convoluted, one simple example that college students experience regularly serves to illustrate the idea. Student feedback on teaching involves students completing a questionnaire about their experience in a particular section of a course. The information collected is not intended to measure anything about the students, but, rather, about the instructor. The instructor is assigned values on a number of measures by aggregating the students' responses. Aggregated means that scores from one unit of analysis are "rolled up" to a "larger" unit of analysis which includes the smaller units. The aggregation might use an averaging or percentages or other calculations, but the point is that the collection of information from students is translated into characteristics assigned to the instructor.

While student feedback on teaching is a good illustration of aggregated self-report, there are many others. If you wanted to study the relationship between community policing strategies and neighborliness, you could conduct surveys of city residents and average their feelings of community by precinct or neighborhood to create scores for neighborliness for distinct areas of the city. You could study the relationship between parental leave policies and job satisfaction by calculating average job satisfaction ratings by company and relate them to the policies of those companies. (The policies might have been measured by representative other-report as described above.) When you are using aggregated self-report you need to be very careful to define your units of analysis. Sense of community may at one level be a characteristic of individuals, but neighborliness as an aggregated sense of community is a characteristic of some larger group such as neighborhood. Furthermore, keep in mind that the use of aggregated self-report presents one of the unusual research situations in which the observational unit (individual residents, for example) is not the same as the unit of analysis, whereas most of the time observational unit and unit of analysis are the same.

Enhancing Response Rates and Monitoring Returns

As mentioned above, response rate refers to the proportion of the sample that actually completes a questionnaire or interview. Anything short of 100% sets up a two-tiered generalizability problem: (1) how well do the collected surveys represent the sample and (2) how well does the sample represent the target population? Low response rates do not necessarily mean that completed surveys do not represent the sample, but the lower the

response rate, the more likely it is that the people who responded differ systematically from those who did not (**non-response bias**). Depending on the amount and accuracy of information in the sampling frame, there are ways to describe non-response bias and, to some extent, take it into account when reporting results. Rather than work through those statistical gymnastics, however, it is better to enhance response rate so as to reduce non-response bias in the first place.

As you know, data collection mode affects response rate. In general, face-to-face interviews yield high response rates. Phone interviews conducted with an affiliated audience (a university's alums, a union's current or past members, members of a congregation, etc.) usually yield fairly high response rates as well. Non-affiliated phone interviews ("cold calls" to randomly selected persons with no connection to the researcher or sponsoring organization) tend to do better than self-administered surveys. That may change as the public becomes more saturated with phone interviews and as telecommunications technology enables more sophisticated call screening. A well-trained interviewer making at least three attempts per case with a good introductory solicitation is the key to enhancing response rate with phone interviews.

The greater challenge to response rate stems from self-administered modes. Hand-in approaches can produce very high response rates as long as adequate time is provided and respondents do not feel that they are being deprived of time for something more valuable when filling out the questionnaire. For example, working with teachers to set aside class time for students to fill out a survey is better than asking students to do it during recess or lunch. If you gain access to a group's meeting time, such as a labor union or retreat attendees' follow-up get-together, it is best to schedule survey administration for the beginning or middle of the meeting (respecting the meeting agenda, of course), rather than put it at the end when the alternatives are (a) completing the survey or (b) going home. Not only do you have to overcome the perception of lost opportunity costs to convince people to stay for the survey, but you also introduce non-response bias because the kinds of people who stay for the survey are almost certainly different from those who go home.

That leaves us with mail-in and web-based. Strategies for administering surveys have been developed by Dillman (1978, 2000). Under the Dillman method, the fundamental requirement is to make sure that the presentation of the invitation to participate is professional and neat—professionally printed envelopes and letterhead stationery, attractive but not fancy type fonts, well-written cover letter, attention to detail in addressing, and so on. As long as resources permit, mail-in surveys should be sent out with "Address Services Requested" so that mail will be forwarded and address changes are noted. Additionally, postage paid return envelopes should be provided. All of these things communicate to your target audience that you, the researcher, take this survey seriously.

In both mail-in and web-based, follow-up reminders are helpful, but the form will vary by mode. A rule of thumb is to count the day of initially sending out the survey or web link as Day One. One week from Day One, a reminder is sent (post card or e-mail), thanking those

who already returned the survey and encouraging those who did not. Three weeks from Day One, another packet is sent, with another cover letter and another survey ("in case the first was misplaced"). The second cover letter should present the same basic information as the first, but in new words. The second packet, if at all possible, should be sent only to non-respondents to that point, which makes it necessary to have some way of monitoring returns. Four weeks from Day One, another post card or e-mail reminder is sent, again only to the non-respondents. With mail-in surveys, a third mailing may have some impact, but it will likely be slight. One possibility is to target the third mailing (whole new packet with new cover letter and another questionnaire) to non-respondents from under-represented groups in the sample. Web-based survey administration follows similar patterns, but (a) the links to the on-line survey are sent in e-mail messages instead of re-mailing paper copies, and (b) the time table can be shorted considerably from a total of four weeks to about ten to fourteen days.

The value of offering incentives is debatable, but worth considering if resources allow and suitable incentives can be identified for the target population. There are three variations on incentives. (1) Include the incentive, usually just a token of appreciation, along with the cover letter and instrument (this can be done with e-mail, as some kind of e-coupon). This kind of incentive is intended to make people feel like they ought to respond, since they have already been given something for doing so. (2) Offer an incentive to all respondents upon submission of a completed survey. Once they return the survey or submit it over the web, a gift of some kind is sent back to them. For hand-in administration, this can be accomplished by offering something (food works well) for those who participate. (3) Have all respondents entered into a drawing for some prize or prizes. If the prize is substantial enough, it is hoped that people will be motivated to complete the survey and get a free chance in a participation lottery.

Of course, resources are a critical consideration here. For many projects, you might want to devote resources to a larger sample or better training rather than to incentives. But two other factors are also important. The first is whether the incentive introduces a response bias. If the incentive is, for example, tickets to a sporting event, non-sports fans might actually be disinclined to participate. If the incentive does not have universal appeal for your sample, it may backfire. Second, even if a neutral and broadly appealing incentive can be found, you have definitely introduced an extrinsic motivation to complete the survey. Having influenced *why* people participate, you might also be influencing *how* they participate. You have to ask yourself, "Might the incentive change what kind of information respondents provide?" If so, the benefit of enhancing response rate should be weighed against the loss of validity.

Finally, some survey administration can be connected to events in the life of the target population in ways that enhance response rate. I call these events "hooks." On one extreme, participation in the event or access to a desired something can be dependent on survey completion. For example, a school could withhold issuing transcripts until a survey is

completed. This option should be used sparingly, since it could easily violate the ethical guideline of voluntary participation. The use of hooks does not need to be that heavy-handed to be effective. If you ask people to fill out a survey at the same time that they gather for an important event (staff picnic, advising nights, informational meetings, etc.), you are taking advantage of the group-administered survey. Alternatively, you could mail people the survey beforehand, but ask them to turn it in at those gatherings. The use of the hook applies only when the target population consists wholly of members of an organization who, as members, have other events or requirements in common.

SAMPLING

From the previous projects you should be familiar with the basics of sampling:

- A **sample** is a subset of some group, called the population, which you want to study. It can be selected in a variety of ways.
- **Sample generalizability** is the extent to which the findings from a specified subset of a population apply to the rest of the population from which the sample was drawn.
- The first step is to define the target population, the set of cases to which you would like to be able to generalize your findings.
- With the target population defined, you determine whether there is a sampling frame available, which is a listing or locating rule of inclusion containing as many elements of the target population as possible.
- Sampling frames might be an assembly of lists, not a single list.
- If a sampling frame exists, a probability sample can, and usually should, be used. If not, nonprobability techniques must be used.
- Because of imperfections in the sampling frame or discrepancies between the frame and the cases actually available at the time of selection, the population from which a sample is drawn is known as the study population.
- Probability techniques rely on the principle of **random selection**. Randomness must meet two criteria: (1) all cases have an equal chance of being selected and (2) the selection of any one case does not affect the chances of selecting any other case.
- There are a number of nonprobability techniques, but the most systematic is referral sampling, in which an initial set of cases are selected based on their known membership in the population, then other cases are selected as referrals from cases already selected.
- Conducting a study on an entire population, as opposed to a sample, is referred to as using a **census**. Census studies can be large or small, depending on the nature of the target population.

At this point in the development of your research skills, having precise terminology for your work is as important as knowing what to do. In particular, you want to describe the parts of your sample clearly. An **element** is a single one of whatever constitutes the population. Most of the time, the element is the same as the sampling unit, but in multi-stage sampling you might have a mix of sampling units. A **sampling unit** is the thing within a set that you can choose. It is easier to illustrate than to define. Suppose your target population is high school athletes, specifically basketball, track, and soccer players because you want equal numbers of males and females. Suppose further that you have decided to limit your study to the state in which you live. Now there is no single list of all the designated athletes in the state, but every high school would have rosters and you could get a list of high schools. Further, you do not have to select students from all high schools to generalize to the whole state, as long as you get a representative sample of high schools. At one point in your sampling procedure, then, high schools are the sampling units. After you have the high schools, the individual athletes on the rosters are your sampling units. Finally, there is the **observation unit**, which is the element which is included in the study and on which you make your observations. Unless you happen to be utilizing **other report** (if, for example, you were asking a coach to answer a set of questions about injury rates for his or her players), the observation unit will be the same as the element.

For this project, we will go into more detail on the kind of advanced probability techniques commonly used in survey research. Even the more advanced techniques, however, are based on one of two basic methods. **Simple random sampling** is like pulling names from a hat. All elements in the frame are available for selection. It could be done manually, which would be cumbersome and time-consuming. Instead, computers can electronically simulate the process. **Systematic random sampling** is the second basic method. Again, you start with all elements in the frame. This method proceeds in three steps. First, determine the proportion of cases that you want to pull (one in 20, one in 13, one in 4, one in 60, whatever). That number is your *interval*. Second, you randomly select one case from the first interval, so that, for example you choose the 26^{th} case from the first 60 cases. Third, you select every 60^{th} cases after that (the 86^{th}, 146^{th}, etc.). Systematic random sampling is useful if the sampling frame has a physical arrangement, such as names in a phonebook, houses in a neighborhood, or records in a file cabinet.

If you could pull an efficient and representative sample with just one round of selections, simple and systematic random techniques would be all you need. Most sampling, though, is more complex than that. Your population might be geographically dispersed, but your resources limit your travel budget. The assembly of an absolutely complete sampling frame might be unwieldy, such as all students playing high school varsity sports in the United States. Your population might consist of many subpopulations with distinct characteristics that you believe will affect your outcomes. These challenges and others can be addressed through stratified and cluster sampling. It is easiest to explain these two strategies together

because they both break populations down into subgroups as part of the random selection process. For a variety of reasons, which should become more obvious after reading this section, you might want to select first some groupings of your elements, and then select individual elements from within the groups in a second or subsequent step.

One kind of subgroup that shares already occurring characteristics is called a **stratum** (the plural is *strata*). Obviously, populations of people can be subdivided by race/ethnicity, marital status, age, grade level, and so on. So, in a population of individuals, some are males and some are females. Some are engineering majors and some are liberal arts majors. In a population of universities, some are private and some are public; furthermore some are community colleges, some are four-year colleges and some are doctoral-granting. In a population of battered women's shelters, some allow long-term stay, others do not. In a population of sex education programs, some are abstinence-based and others are contraception-based. Each of the subgroups is a stratum.

The other kind of grouping is a **cluster**, and it is defined by geographical location. Types of clusters include states, counties, cities, apartment buildings, neighborhoods, residence halls, etc. As will be explained shortly, one cluster should not be systematically different from another.

In the case of both clusters and strata, to be useful in sampling, an element must belong to one and only one group. In other words, you cannot use strata if membership in one stratum could overlap with membership in another stratum. You could not, for example, stratify high school students by the sport they play, since many students play more than one sport.

The difference between strata and clusters is not merely the kind of grouping it is—that is, whether it is based on a characteristic or location. In fact, for the purpose of sampling design, that is relatively unimportant. The really important difference is how the groupings compare to each other and how the elements within the group compare to each other. To use **stratified sampling**, the elements should be mostly similar to each other within the stratum, but mostly different from elements in other strata. Conversely, to use **cluster sampling** you want the elements to be mostly different from each other within a cluster, but any given cluster to be mostly similar to other clusters. Another way of saying this is that strata are *homogeneous within and heterogeneous between,* while clusters are *heterogeneous within and homogeneous between*.

Let us examine why this is so important. The goal of stratified random sampling is to select elements from within a stratum that represent the stratum as a whole. You make sure that all the strata are represented so that differences within the entire population are captured. Suppose you wanted to study how much time students spend studying each week. Just for the sake of example (these numbers are fictitious and not meant to impugn any major), suppose all underwater hotel management majors studied 8 hours per week, that all space travel management majors studied 19 hours per week, and that all interspecies communications majors studied 25 hours per week. Under these

circumstances, you really would only need to sample one person from each major, because each of those three people perfectly represented his or her stratum. You could take a very small sample and achieve generalizability. Now, real life is not that simple, but when the condition of homogeneous within and heterogeneous between is met, stratified sampling makes good sense.

What about clusters? Clusters are typically used when your target population is geographically spread out and you lack the wherewithal to collect data from elements here, there, and everywhere. Moreover, you might not have a single sampling frame that includes the entire list of elements in your target population, just because of the size or geographical spread of the population. But you could get a sampling frame within a cluster. The goal of cluster sampling is to select some groupings within the population, then elements within those clusters because you are fairly certain that the diverse elements within Cluster A, which you did select, are as representative of the whole target population as are elements within Clusters B, C, and D, which you did not select. You can then concentrate your data collection efforts in a more confined area without jeopardizing generalizability to the target population. Of course, you strengthen your sample by selecting an appropriate number of clusters, not just one. Furthermore, clustering is often done in multiple stages, as in selecting regions within a country, then states within the regions, then counties with the state, then census tracts or cities within the county. Keep in mind, however, how your definition and selection of clusters affects the extent to which your study population reflects your target population. If you cluster a college campus by residence hall, you might have to admit that your study population represents only on-campus students, not the entire student body.

The use of stratified and cluster sampling puts special demands on the identification of sampling frames. For stratified sampling, you need to make sure that the sampling frame of the whole target population includes information about the stratum to which each element belongs. If not, you cannot sort the elements into strata in order to select randomly from within each stratum. If, for instance, you want to stratify by marital status, your sampling frame has to have marital status for each person along with the usual contact information. If you are doing multi-stage stratified sampling, information on all strata characteristics has to be in the sampling frame. For cluster sampling, you actually have to generate a sampling frame for each stage of clustering. A list of states would not be a problem, nor counties within each state. But it might take special efforts to locate a full sampling frame for smaller clusters like apartment complexes or census tracts. Keep in mind also that, once you get to the final cluster stage, you need to generate a list of the sampling units within that cluster. Your dependence on so many sampling frames introduces the possibility of errors at each stage, which is why cluster sampling requires larger final sample size than non-cluster techniques. Larger samples counteract the multiple potential error sources.

Two additional strategies can make probability sampling more efficient and more generalizable. One is the use of **disproportionate stratified sampling**. If the sizes of the

strata into which the population is divided are decidedly unequal, with some strata much smaller than others, you probably want to take a larger *proportion* of cases from the small strata than you do from the large strata. This is frequently done when stratifying by race. You might select something like 60% of Native Americans, 30% of Asian Americans, African Americans, and Latinos, but only 5% of whites. The exact proportions will vary from study to study, but the use of unequal proportions helps to make sure that you have enough elements from the small strata to perform statistical analyses. If the strata you use are all about the same size in the target population, proportionate stratified sampling will suffice.

You might also use disproportionate stratified sampling when you want to weight your sample to over-represent cases that are more relevant to your research question. This is not necessarily a way to achieve optimum generalizability, but does help when the purpose of the study is to inform policy decisions. If you are studying options for restructuring teacher compensation, for example, you might oversample teachers who have more seniority or teachers who are heads of households.

The other useful strategy is combinations of clustering and stratifying. Once you have selected your final cluster, for example, you may need to stratify by some characteristic like type of school or rural-suburban-urban setting. It is also useful sometimes to stratify the clusters. For example, you may want to make sure that all regions of the country are properly represented in your final sample, so you divide the country into regions and perhaps use disproportionate stratified sampling to select states from within the regions.

Finally, keep in mind that, with both cluster and stratified sampling, randomness is always the principle for selection at any stage. At the last stage, once you have a sampling frame for the final clusters or strata, either simple random or systematic random sampling is used to select the elements of your sample.

There is one more option among sampling strategies that can be very handy with obscure or hard to identify populations. That is **over-sampling**, which carries most of the advantages of probability sampling into what might otherwise be a nonprobability population. Over-sampling is like casting a wide net into the sea to collect study samples of a certain kind of fish that you know is rare and dispersed. From within the total catch, you save the rare fish and return the others to the sea. Your sea is the larger population which includes some unknown number of the rare cases you want to study. Your net is the over-sampling technique.

For example, there is no sampling frame of women who have experienced acquaintance rape nor one of people with eating disorders. But if you sampled college students in general, the chances are pretty good that some of them will be women who have experienced acquaintance rape or are people with an eating disorder. With appropriately sensitive wording and well-thought filter questions (which determine whether the respondent or interviewee fits the description of your target population), you may be able to create a large enough sample to conduct your study. You might not find the remaining surveys very useful,

and for that reason, you would want to make sure that you had not inordinately imposed on their time (in other words, the survey was not too long).

An example of using over-sampling or multi-stage sampling to identify a hidden population is provided by the work of Jayne Mooney (2000) who conducted one of the largest surveys of domestic violence in Britain. She actually used three stages of sampling to locate potential interviewees. In the first stage, face-to-face interviews were conducted of 1,205 households in North London. The interviewees included both men and women at that stage. Among the women interviewees, those whose partners were not present at the time of the interview were given a paper survey and return envelope. With an 80% response rate to the second survey, which insured privacy although not anonymity (the paper surveys were number-coded to correspond with the face-to-face interviews), the researcher identified a sufficient number of women who had experienced domestic violence. Mooney followed up with those women for in-depth interviews.

Sample Size. There is no ideal sample size, but a few basic guidelines help determine what size is right for a given project. In general, the larger the sample, the better, because a statistical anomaly called sampling error is reduced by quantity alone. That said, however, representativeness can be enhanced without having to sample to the bursting point. Because most statistical procedures use the size of the sample in its square root form in the denominator of computations, an increase in sample size from, say, 500 to 1000 improves accuracy quite a bit, but, ironically, going from 5,000 to 10,000 improves accuracy only slightly. Another thing to keep in mind is that the size of the *target population* has no *direct* importance in determining appropriate sample size because large populations can be accurately represented by relatively small samples. Good use of clustering and stratification, sufficient foreknowledge of the population characteristics, and efficient execution of data collection all make it possible to generalize to very large populations with relatively small samples. Surveys generalizable to the adult US population on many economic and political issues, for example, can be fairly stable and accurate based on samples of about 1,750. Some studies do include lots of cases—30,000 and higher—but that is usually because the researchers want to do many breakdowns in the analysis or because they expect to lose many initial participants in later rounds of longitudinal data collection.

Given that larger is generally better, but assuming that the researcher has limited resources to use as wisely as possible, there are five general rules to guide decisions about sample size. These rules apply to probability samples; nonprobability samples tend to be fairly small to begin with and are constrained by mostly indeterminate factors like the saturation point and the extent to which expertise supporting a purposive sample matches the actual characteristics of the target population. With regard to probability samples, however, the following guidelines provide some direction.

- The more diversity there is in the target population on the factors that matter to your study, the larger your sample must be. Not only does the sample have to capture more

cases to represent the population's variety, but statistical estimates lose precision with larger standard deviations. Standard deviations indicate amount of dispersion around the mean which is mathematical lingo for diversity.

- The more precision you want, the larger your sample must be. If it is important that the findings reduce guesswork and wiggle room, then a large sample reduces the confidence interval around all estimates of population characteristics. Smaller confidence intervals allow for more precision in the statistics.
- The more breakdowns you plan to do, the larger your sample must be. If you want to be able to report that the average person in your population has X amount of something, you can get by with a smaller sample. But if you want to report that left-handed, blue-eyed, popcorn-lovers born in May have Y amount of that something while right-handed, hazel-eyed, popcorn-haters born in August have Z amount of it, then you need a larger sample.
- The type of sample you draw also affects sample size. Systematic random samples need to be larger than simple random samples because of the slight possibility of periodicity problems (see Project One). Cluster samples need to be larger (sometimes much larger) than simple random or stratified samples because of the introduction of sampling error at each clustering stage. Stratified samples, if using strata that are very internally homogeneous, can be relatively smaller than simple or systematic random samples.
- To make sure that you have enough cases for statistical analysis, a good rule of thumb is to have 20 cases per cell. A cell is the smallest breakdown you expect to analyze; if, for example, you know that you want breakdowns by gender (male and female), age group (six age groups) and race/ethnicity (eight groups), you have 2 X 6 X 8 cells (96) and the smallest of these cells should have at least 20 cases. For survey research, keep in mind that this means you need 20 respondents per cell, not just 20 of each in your sample. On the high end, keep in mind that a cell larger than 140 is fine, but confidence intervals for Ns 140 and above are about the same.

RELATIONSHIPS

In Project Three, you learned that the richly detailed explanation afforded by field research lends that strategy to **idiographic causality**. That type of relationship description links events to one another in causal chains and inter-connections. It also typically is limited in generalizability because it is very situation specific. Surveys, in contrast, are excellent tools for uncovering **nomothetic causality**. A nomothetic explanation for racial disturbances or riots, for example, would emphasize the ways that riots resemble other forms of collective behavior such as the nature of grievances and precipitating events. An idiographic explanation for a particular riot would identify conditions in the area where the riot occurred, particularly how certain events were followed by other events and how leaders

and crowds reacted to those events in ways that lead to the initial outbreak of violence and how the initial outbreak grew into wider scale violence.

Nomothetic explanations illuminate the broad sweep of how events and characteristics tend to affect other factors. When we make nomothetic claims, we know that exceptions occur, but that the general truth of the claims will hold under most situations. To reach these conclusions, we must look for patterns in the way that change in one set of conditions is related to changes in other phenomena, while narrowing the plausible explanations down to the ones that have the fewest exceptions. Because surveys can reach large samples and tend to generate quantitative measures for multiple concepts, they are well suited for nomothetic explanations.

For your data collection for Project One you are interviewing your small panel of students repeatedly, thus allowing you to establish **time order** of some of your data points. Time order, in turn, allows you to establish causality. Longitudinal designs are the primary, but not only, way that surveys establish time order. For this project, you will collect all of the information at a single point in time, making this a **cross-sectional study**. That does not mean that you cannot establish causality, but it does make it more challenging. For one thing, certain traits are ascribed, meaning that they are set at birth (gender and ethnicity are obvious), and can therefore be construed to precede the development of personality traits and other life experiences. In addition, and more importantly, time order can be captured in the type of questions. For example, you can ask graduating seniors what kinds of activities they did while in school and also ask how they think they have changed or how much they learned since their first year. It can be argued, then, that the activities were among the factors that contributed to those changes. Another way to establish time order in a cross-sectional design is to ask some questions about events in a person's life, such as the birth of a child, marriage, earning a degree, getting or losing a job, and so on. If other items on the survey operationalize choices or positions that were most likely made or taken after those events, it is reasonable to posit the life events as causes and the choices as effects. Approaches such as these, as you might guess, do not yield robust causal claims and should, therefore, be used sparingly.

Given the cross-sectional nature of this project, we will focus on the ways that surveys help to establish one of the other necessary criteria for establishing causality: **non-spuriousness**, the ruling out of extraneous factors and plausible alternative explanations that might better explain the observed association between what you have hypothesized as cause and effect. Because of the ease with which a survey can operationalize multiple concepts, it can include a host of control variables. **Control variables** are factors that you can take into account when analyzing relationships so that you are, in effect, comparing only cases that are similar on the controls when you look at the hypothesized cause and effect.

We know, for example, that average educational attainment among minorities is lower than it is among whites. We also know that education increases income level. So if we control for education level while examining the relationship between race and income, we

can conclude that, controlling for differences in years of schooling, minorities earn less than whites. That is to say, comparing whites and minorities at the same level of education, whites make more than minorities. This is a relatively simple example. There are, in fact, many other factors that have an effect on income and that also differ systematically by ethnicity.

Another way to rule out the influence of certain factors is to hold them **constant**, which is what we are doing when we limit the target population to people who share characteristics. In survey research, that is the only way to hold something constant and is also referred to as **bounding the study**. You already know that results from a sample from a defined target population cannot be generalized to other populations. But it also has an effect on what we can say about relationships within the study. If a sample is drawn entirely from, say, adults with children, we are ruling out parenthood as a factor in any other relationships we observe. That is, if we find that the respondents in our study were highly religious, we cannot claim that having children makes people more religious, because we cannot compare our respondents with people who do not have children. Likewise, if a sample is drawn entirely from people of a single religious affiliation, religious affiliation cannot be used to explain any relationships we observe. That is not to say that religious affiliation is, in the end, not an explanatory factor, because it might be. It's simply that our study could not make the claim.

Like bounding a study, response rates have an effect on both generalizability and relational validity. Recall from above that there are two types of response rates. First, some individuals in the sample return the survey and some do not—allowing us to calculate the **overall response rate**. Second, some respondents answer a particular item on the questionnaire and others do not—allowing us to derive the **item response rate.** The real threat to validity arises when the kind of people who respond are systematically different from the kind who do not respond, which creates **non-response bias**. Overall non-response bias undermines generalizability; item non-response bias undermines relational validity. Suppose for example, that a substantial proportion of your respondents skip a question about household income (this, in fact, happens often). If you hypothesize that income affects something else like social network complexity, not only do you have to contend with possible spurious factors associated with income, but you also have the problem that you lost a large number of cases from your measure of the association between income and social network complexity. If you have sufficient controls, this disadvantage can be partially overcome, but it still undermines the strength of your causal claims. Of course, you cannot force respondents to answer every question, but you can design surveys so that sensitive questions are worded such that they encourage response and you can emphasize the confidentiality of the results. Moreover, as mentioned, good control measures help reduce the impact of item non-response bias.

ETHICS

The ethical concerns for self-administered surveys differ very little from those for interviews. Again, the main ethical principles to be maintained are **informed consent** (with its parallel principle of **non-coercion**) and **protection of privacy**. Cover letters (paper or electronic) have to disclose the sponsorship and purposes of the survey, the intended uses of the data, and procedures for protecting confidentiality. As long as the cover letter makes specific mention of the implied consent, it is permissible to construe submission of the survey as consenting to participate. As mentioned above, group-administered or hand-in surveys have to be administered in ways that people do not feel forced to participate.

CONNECTIONS TO OTHER PROJECTS

The researcher can design a survey that shares some of the characteristics of experimental design and has, therefore, enhanced relational validity strengths. The **split ballot design** and its cousin strategy the **factorial survey**, have two or more versions of what are, for the most part, parallel instruments. The main difference between split ballot designs and factorial surveys is that the former intends to understand some specific feature of survey design, while the latter manipulates an independent variable just like any true experiment would. We will describe the split ballot design here and factorial surveys along with the experiment project.

One of the most interesting facts about research methods is this: much of what we have learned about administering *surveys* comes from doing *experiments*. If you administer a survey in two slightly different forms to randomly assigned groups (half the sample gets version A and the other half gets version B), you can learn a great deal about the effects of that slight difference. This type of research involves manipulating a characteristic of the survey (administration or instrument), and observing any number of outcomes: typically either response rates or the way respondents answer questions. For example, you could manipulate the way the survey looks without changing its content (e.g., use of color, fancy fonts, graphics, amount of white space, etc.), then see whether you get a higher response rate with one look over the other one. You could examine the effects of including a neutral category in response options to see which way (positive or negative) respondents are likely to shift if given forced-choice options. Researchers have examined the effect of different kinds of envelopes on mail-in response rates. Recently, many studies have compared a paper version of a survey with its on-line counterpart. People have studied the effects of incentives, leading questions, item placement, funnel sequences, and numerous other features of survey design. I once experimented with slightly different wording of agree-disagree statements for a scale I was constructing (version A: “When I see someone in need, I am usually moved to help.” Version B: “When I see a stranger in need, I am usually moved

to help."). I found that people are far less likely to agree with the latter statement than the former, and that those who agree with the latter statement tend to score higher on other items in a social responsibility scale.

KEY TERMS

Aggregated self-report
Bounding a study
Census
Closed-ended question
Cluster
Cluster sampling
Cognitive interview
Combinative scale
Construct
Construct validity
Contingent question
Control variable *
Criterion validity (predictive and concurrent)
Cross-sectional time design
Data collection mode
Disproportionate stratified sampling
Double-barreled wording
Element *
Email survey
Exhaustive and mutually exclusive
Factorial survey
Familiarity other-report
Filter question
Forced-choice question
Funnel sequence
Group-administered survey
Hand-in survey
Idiographic causality
Idiosyncratic variation
Index or scale
Indicator weighting
Instrument
Instrument construction
Inter-item correlation
Interpretive question
Interview *
Inverted funnel sequence
Item neutrality
Item response rate
Item singularity
Item-scale correlation
Iterative scale
Leading question
Likert scale
Loaded wording
Mail-in survey
Nomothetic causality
Non-response bias
Non-spuriousness
Observation unit *
Omnibus survey
Open-ended question
Other-report
Over-sampling
Questionnaire *
Random selection
Representative other-report
Respondent *
Response categories/options
Response rate (full and partial)
Reverse coding
Rules of combination
Sampling unit
Self-administered survey *
Semantic differential scale
Simple random sampling
Skip logic
Skip pattern
Split ballot design
Stratified sampling
Stratum
Study population
Systematic random sampling
Target population
Web-based survey

* Reinforced from a previous exercise or project