15 Surveys & Instruments

Shortly after the terrorist attacks in New York City and Washington, DC, in September of 2001, researcher (Lerner, Gonzalez, Small & Fischhoff, 2003) reported an Internet-based survey of nearly 2,000 American teens and adults ranging in age from 13 to 88. They asked participants about their reactions to the attacks and for their judgments of various terrorism-related and other risks. Among the results were that the participants tended to overestimate most risks, that females did so more than males, and that there were no differences between teens and adults. The most interesting result, however, had to do with the fact that some participants were "primed" to feel anger by asking them what made them angry about the attacks and by presenting them with a photograph and audio clip intended to evoke anger. Others were primed to feel fear by asking them what made them fearful about the attacks and by presenting them with a photograph and audio clip intended to evoke fear. As the researchers hypothesized, the participants who were primed to feel anger perceived less risk than the participants who had been primed to feel fear—showing how risk perceptions are strongly tied to specific emotions.

The study by Lerner and her colleagues is an example of survey research in psychology—the topic of this chapter. We begin with an overview of survey research, including its definition, some history, and a bit about who conducts

it and why. We then look at survey responding as a psychological process and the implications of this for constructing good survey questionnaires. Finally, we consider some issues related to actually conducting survey research, including sampling the participants and collecting the data.

Learning Objectives

1. Define what survey research is, including its two important characteristics.
2. Describe several different ways that survey research can be used and give some examples.
3. Explain what a **context effect** is and give some examples.
4. Define sampling bias in general and non-response bias in particular. List some techniques that can be used to increase the response rate and reduce non-response bias.
5. Define **instrument reliability**, including the different types and how they are assessed.
6. Define **instrument validity**, including the different types and how they are assessed.

What Is Survey Research?

Survey research is a quantitative and qualitative method with two important characteristics. First, the variables of interest are measured using self-reports (using questionnaires or interviews). In essence, survey researchers ask their participants (who are often called respondents in survey research) to report directly on their own thoughts, feelings, and behaviors. Second, considerable attention is paid to the issue of sampling. In particular, survey researchers have a strong preference for large random samples because they provide the most accurate estimates of what is true in the population. In fact, survey research may be the only approach in psychology in which random sampling is routinely used. Beyond these two characteristics, almost anything goes in survey research. Surveys can be long or short. They can be conducted in person, by telephone, through the mail, or over the Internet. They can be about voting intentions, consumer preferences, social attitudes, health, or anything else that it is possible to ask people about and receive meaningful answers. Although survey data are often analyzed using statistics, there are many questions that lend themselves to more qualitative analysis.

Most survey research is non-experimental. It is used to describe single variables (e.g., the percentage of voters who prefer one presidential candidate or another, the prevalence of schizophrenia in the general population, etc.) and also to assess statistical relationships between variables (e.g., the relationship between income and health). But surveys can also be used within experimental research. The study by Lerner and her colleagues is a good example. Their use of self-report measures and a large national sample identifies their work as survey research. But their manipulation of an independent variable (anger vs. fear) to assess its effect on a dependent variable (risk judgments) also identifies their work as experimental.

History and Uses of Survey Research

Survey research may have its roots in English and American "social surveys"

conducted around the turn of the 20th century by researchers and reformers who wanted to document the extent of social problems such as poverty (Converse, 1987). By the 1930s, the US government was conducting surveys to document economic and social conditions in the country. The need to draw conclusions about the entire population helped spur advances in sampling procedures. At about the same time, several researchers who had already made a name for themselves in market research, studying consumer preferences for American businesses, turned their attention to election polling. A watershed event was the presidential election of 1936 between Alf Landon and Franklin Roosevelt. A magazine called Literary Digest conducted a survey by sending ballots (which were also subscription requests) to millions of Americans. Based on this "straw poll," the editors predicted that Landon would win in a landslide. At the same time, the new pollsters were using scientific methods with much smaller samples to predict just the opposite—that Roosevelt would win in a landslide. In fact, one of them, George Gallup, publicly criticized the methods of Literary Digest before the election and all but guaranteed that his prediction would be correct. And of course, it was, demonstrating the effectiveness of careful survey methodology (We will consider the reasons that Gallup was right later in this chapter). Gallup's demonstration of the power of careful survey methods led later researchers to local, and in 1948, the first national election survey by the Survey Research Center at the University of Michigan. This work eventually became the American National Election Studies (https://electionstudies.org/) as a collaboration of Stanford University and the University of Michigan, and these studies continue today.

From market research and election polling, survey research made its way into several academic fields, including political science, sociology, and public health—where it continues to be one of the primary approaches to collecting new data. Beginning in the 1930s, psychologists made important advances in questionnaire design, including techniques that are still used today, such as the Likert scale (defined below). Survey research has a strong historical association with the social psychological study of attitudes, stereotypes, and prejudice. Early attitude researchers were also among the first psychologists

to seek larger and more diverse samples than the convenience samples of university students that were routinely used in psychology (and still are).

Survey research continues to be important in psychology today. For example, survey data have been instrumental in estimating the prevalence of various mental disorders and identifying statistical relationships among those disorders and with various other factors. The National Comorbidity Survey is a large-scale mental health survey conducted in the United States (see http://www.hcp.med.harvard.edu/ncs). In just one part of this survey, nearly 10,000 adults were given a structured mental health interview in their homes in 2002 and 2003. The table below presents results on the lifetime prevalence of some anxiety, mood, and substance use disorders. Obviously, this kind of information can be of great use both to basic researchers seeking to understand the causes and correlates of mental disorders as well as to clinicians and policymakers who need to understand exactly how common these disorders are.

Table of Some Lifetime Prevalence Results from the National Comorbidity Survey. Note that the lifetime prevalence of a disorder is the percentage of people in the population that develop that disorder at any time in their lives.

Lifetime prevalence			
Disorder	**Total**	**Female**	**Male**
Generalized anxiety disorder	5.7	7.1	4.2
Obsessive-compulsive disorder	2.3	3.1	1.6
Major depressive disorder	16.9	20.2	13.2
Bipolar disorder	4.4	4.5	4.3
Alcohol abuse	13.2	7.5	19.6
Drug abuse	8.0	4.8	11.6

And as the opening example makes clear, survey research can even be used as a data collection method within experimental research to test specific hypotheses about causal relationships between variables. Such studies, when conducted on large and diverse samples, can be a useful supplement to laboratory studies conducted on university students. Survey research is thus a flexible approach that can be used to study a variety of basic and applied

research questions.

Constructing Surveys

The heart of any survey research project is the survey itself. Although it is easy to think of interesting questions to ask people, constructing a good survey is not easy at all. The problem is that the answers people give can be influenced in unintended ways by the wording of the items, the order of the items, the response options provided, and many other factors. At best, these influences add noise to the data. At worst, they result in systematic biases and misleading results. In this section, therefore, we consider some principles for constructing surveys to minimize these unintended effects and thereby maximize the reliability and validity of respondents' answers.

Before looking at specific principles of survey construction, it will help to consider survey responding as a psychological process.

Consider, for example, the following questionnaire item:

How many alcoholic drinks do you consume in a typical day?

_____ *a lot more than average*

_____ *somewhat more than average*

_____ *average*

_____ *somewhat fewer than average*

_____ *a lot fewer than average*

Although this item at first seems straightforward, it poses several difficulties for respondents. First, they must interpret the question. For example, they must decide whether "alcoholic drinks" include beer and wine (as opposed to just hard liquor) and whether a "typical day" is a typical weekday, typical weekend day, or both. Even though Chang and Krosnick (2003) found that asking about "typical" behavior has been shown to be more valid than

asking about "past" behavior, their study compared "typical week" to "past week" and may be different when considering typical weekdays or weekend days). Once respondents have interpreted the question, they must retrieve relevant information from memory to answer it. But what information should they retrieve, and how should they go about retrieving it? They might think vaguely about some recent occasions on which they drank alcohol, they might carefully try to recall and count the number of alcoholic drinks they consumed last week, or they might retrieve some existing beliefs that they have about themselves (e.g., "I am not much of a drinker"). Then they must use this information to arrive at a tentative judgment about how many alcoholic drinks they consume in a typical day. For example, this mental calculation might mean dividing the number of alcoholic drinks they consumed last week by seven to come up with an average number per day. Then they must format this tentative answer in terms of the response options actually provided. In this case, the options pose additional problems of interpretation. For example, what does "average" mean, and what would count as "somewhat more" than average? Finally, they must decide whether they want to report the response they have come up with or whether they want to edit it in some way. For example, if they believe that they drink a lot more than average, they might not want to report that for fear of looking bad in the eyes of the researcher, so instead, they may opt to select the "somewhat more than average" response option.

From this perspective, what at first appears to be a simple matter of asking people how much they drink (and receiving a straightforward answer from them) turns out to be much more complex. Measurement of alcohol use is a good example of where a stronger methodological approach is to use an established "instrument" for which existing studies of reliability and validity are available. This may also highlight important theoretical questions such as the difference in alcohol use scales aimed an quantifying behavior versus those specifically aimed at identifying possible abuse or addiction disorders (e.g., Greenfield, 2000)

Context Effects on Survey Responses

Again, this complexity can lead to unintended influences on respondents' answers. These are often referred to as context effects because they are not related to the content of the item but to the context in which the item appears (Schwarz & Strack, 1990).For example, there is an item-order effect when the order in which the items are presented affects people's responses. One item can change how participants interpret a later item or change the information that they retrieve to respond to later items. Strack, Martin & Schwarz (1988) asked college students about both their general life satisfaction and their dating frequency. When the life satisfaction item came first, the correlation between the two was only −.12 (very weak), suggesting that the two variables are only weakly related. But when the dating frequency item came first, the correlation between the two was +.66 (strongly correlated), suggesting that those who date more have a strong tendency to be more satisfied with their lives. Reporting the dating frequency first made that information more accessible in memory so that they were more likely to base their life satisfaction rating on it.

The response options provided can also have unintended effects on people's responses (Schwarz, 1999). For example, when people are asked how often they are "really irritated" and given response options ranging from "less than once a year" to "more than once a month," they tend to think of major irritations and report being irritated infrequently. But when they are given response options ranging from "less than once a day" to "several times a month," they tend to think of minor irritations and report being irritated frequently. People also tend to assume that middle response options represent what is normal or typical. So, if they think of themselves as normal or typical, they tend to choose middle response options. For example, people are likely to report watching more television when the response options are centered on a middle option of 4 hours than when centered on a middle option of 2 hours. To mitigate against order effects, rotate questions and response items when there is no natural order. Counterbalancing or randomizing the order of presentation of the questions in online surveys are good practices for survey

questions and can reduce response order effects that show that among undecided voters, the first candidate listed in a ballot receives a 2.5% boost simply by virtue of being listed first!

Writing Survey Items

Questionnaire items can be either open-ended or closed-ended. Open-ended items simply ask a question and allow participants to answer in whatever way they choose. The following are examples of open-ended questionnaire items.

- *"What is the most important thing to teach children to prepare them for life?"*
- *"Please describe a time when you were discriminated against because of your age."*
- *"Is there anything else you would like to tell us about?"*

Open-ended items are useful when researchers do not know how participants might respond or when they want to avoid influencing their responses. Open-ended items tend to be used in qualitative research or in the early stages of a research project.

Closed-ended items ask a question and provide a set of response options for participants to choose from. The alcohol item just mentioned is an example, as are the following:

How old are you?

_____ *Under 18*

_____ *18 to 34*

_____ *35 to 49*

_____ *50 to 70*

_____ *Over 70*

On a scale of 0 (no pain at all) to 10 (worst pain ever experienced), how

much pain are you in right now?

Have you ever in your adult life been depressed for a period of 2 weeks or more? Yes No

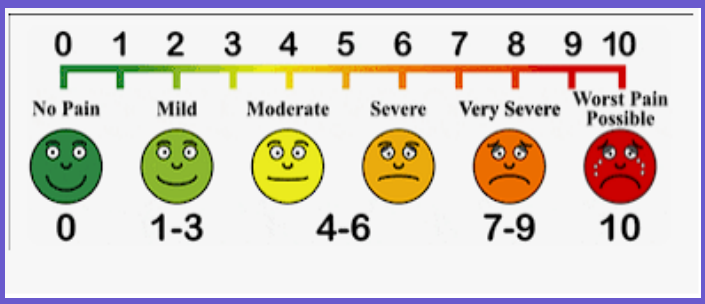
Closed-ended items are used when researchers have a good idea of the different responses that participants might make. They are quantitative in nature, so they are also used when researchers are interested in a well-defined variable or construct such as participants' level of agreement with some statement, perceptions of risk, or frequency of a particular behavior. Closed-ended items are more difficult to write because they must include an appropriate set of response options. However, they are relatively quick and easy for participants to complete. They are also much easier for researchers to analyze because the responses can be easily converted to numbers and entered into a spreadsheet. For these reasons, closed-ended items are much more common.

All closed-ended items include a set of response options from which a participant must choose. For categorical variables like sex, race, or political party preference, the categories are usually listed and participants choose the one (or ones) to which they belong. For quantitative variables, a rating scale is typically provided. A rating scale is an ordered set of responses that participants must choose from. Figure 7.2 shows several examples. The number of response options on a typical rating scale ranges from three to 11—although five and seven are probably most common. Five-point scales are best for unipolar scales where only one construct is tested, such as frequency (Never, Rarely, Sometimes, Often, Always). Seven-point scales are best for bipolar scales where there is a dichotomous spectrum, such as liking (Like very much, Like somewhat, Like slightly, Neither like nor dislike, Dislike slightly, Dislike somewhat, Dislike very much). For bipolar questions, it is useful to offer an earlier question that branches them into an area of the scale; if asking about liking ice cream, first ask "Do you generally like or dislike ice cream?" Once the respondent chooses like or dislike, refine it by offering them relevant choices from the seven-point scale. Branching improves both reliability and validity (Krosnick & Berent, 1993). Although you

often see scales with numerical labels, it is best to only present verbal labels to the respondents but convert them to numerical values in the analyses. Avoid partial labels or length or overly specific labels. In some cases, the verbal labels can be supplemented with (or even replaced by) meaningful graphics.

Example Scales

Here are a few examples of scales with more creative use of graphics or layout used for different kinds of survey responses.



Statement
Academic detailing is a useful form of education that aligns providers' prescribing behavior with evidence-based practice.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

How satisfied are you with *

	Very Unsatisfied	Unsatisfied	Neutral	Satisfied	Very Satisfied
Purchase	○	○	○	○	✓
Service	○	○	○	✓	○
Company Overall	○	○	○	○	✓

1.) On a scale of „no itch" (left) to „worst imaginable itch" (right), how was

Please mark a position between 0 and 10 that best represents your itch with a cross on the line below.

…your itch, on average, in the past 24 hours? 0 —— 10

…your worst itch in the past 24 hours? ——

Likert Scales

In reading about psychological research, you are likely to encounter the term Likert scale. Although this term is sometimes used to refer to almost any rating scale (e.g., a 0-to-10 life satisfaction scale), it has a much more precise meaning.

In the 1930s, researcher Rensis Likert (pronounced LICK-ert) created a new approach for measuring people's attitudes (Likert, 1932). It involves presenting people with several statements—including both favorable and unfavorable statements—about some person, group, or idea. Respondents then express their agreement or disagreement with each statement on a 5-point scale: Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, Strongly Disagree. Numbers are assigned to each response and then summed across all items to produce a score representing the attitude toward the person, group, or idea. For items that are phrased in an opposite direction (e.g., negatively worded statements instead of positively worded statements), reverse coding is used so that the numerical scoring of statements also runs in the opposite direction. The entire set of items came to be called a Likert scale.

Thus, unless you are measuring people's attitude toward something by assessing their level of agreement with several statements about it, it is best to avoid calling it a Likert scale. You are probably just using a "rating scale."

Writing Effective Items

We can now consider some principles of writing questionnaire items that minimize unintended context effects and maximize the reliability and validity of participants' responses. A rough guideline for writing questionnaire items is provided by the **BRUSO** model (Peterson, 2000). An acronym, BRUSO stands for "brief," "relevant," "unambiguous," "specific," and "objective." Effective questionnaire items are brief and to the point. They avoid long, overly technical, or unnecessary words. This brevity makes them easier

for respondents to understand and faster for them to complete. Effective questionnaire items are also relevant to the research question. If a respondent's sexual orientation, marital status, or income is not relevant, then items on them should probably not be included. Again, this makes the questionnaire faster to complete, but it also avoids annoying respondents with what they will rightly perceive as irrelevant or even "nosy" questions. Effective questionnaire items are also unambiguous; they can be interpreted in only one way. Part of the problem with the alcohol item presented earlier in this section is that different respondents might have different ideas about what constitutes "an alcoholic drink" or "a typical day." Effective questionnaire items are also specific so that it is clear to respondents what their response should be about and clear to researchers what it is about. A common problem here is closed-ended items that are "double barreled." They ask about two conceptually separate issues but allow only one response. For example, "Please rate the extent to which you have been feeling anxious and depressed." This item should probably be split into two separate items—one about anxiety and one about depression. Finally, effective questionnaire items are objective in the sense that they do not reveal the researcher's own opinions or lead participants to answer in a particular way. The table below shows some examples of poor and effective questionnaire items based on the BRUSO criteria. The best way to know how people interpret the wording of the question is to conduct a pilot test and ask a few people to explain how they interpreted the question.

Table: BRUSO Model of Writing Effective Questionnaire Items with Examples		
Criterion	Poor	Effective
B—Brief	"Are you now or have you ever been the possessor of a firearm?"	"Have you ever owned a gun?"
R—Relevant	"What is your sexual orientation?"	Do not include this item unless it is clearly relevant to the research
U—Unambiguous	"Are you a gun person?"	"Do you currently own a gun?"
S—Specific	"How much have you read about the new gun control measure and sales tax?"	"How much have you read about the new sales tax?"
O—Objective	"How much do you support the new gun control measure?"	"What is your view of the new gun control measure?"

For closed-ended items, it is also important to create an appropriate response scale. For categorical variables, the categories presented should generally be mutually exclusive and exhaustive. Mutually exclusive categories do not overlap. For a religion item, for example, the categories of Christian and Catholic are not mutually exclusive but Protestant and Catholic are mutually exclusive. Exhaustive categories cover all possible responses. Although Protestant and Catholic are mutually exclusive, they are not exhaustive because there are many other religious categories that a respondent might select: Jewish, Hindu, Buddhist, and so on. In many cases, it is not feasible to include every possible category, in which case an Other category, with a space for the respondent to fill in a more specific response, is a good solution. If respondents could belong to more than one category (e.g., race), they should be instructed to choose all categories that apply.

For rating scales, five or seven response options generally allow about as much precision as respondents are capable of. However, numerical scales with more options can sometimes be appropriate. For dimensions such as attractiveness, pain, and likelihood, a 0-to-10 scale will be familiar to many

respondents and easy for them to use. Regardless of the number of response options, the most extreme ones should generally be "balanced" around a neutral or modal midpoint. An example of an unbalanced rating scale measuring perceived likelihood might look like this:

Unlikely | *Somewhat Likely* | *Likely* | *Very Likely* | *Extremely Likely*

A balanced version might look like this:

Extremely Unlikely | *Somewhat Unlikely* | *As Likely as Not* | *Somewhat Likely* | *Extremely Likely*

Note, however, that a middle or neutral response option does not have to be included. Researchers sometimes choose to leave it out because they want to encourage respondents to think more deeply about their response and not simply choose the middle option by default. However, including middle alternatives on bipolar dimensions can be used to allow people to choose an option that is neither.

Formatting the Survey

Writing effective items is only one part of constructing a survey. For one thing, every survey should have a written or spoken introduction that serves two basic functions (Peterson, 2000). One is to encourage respondents to participate in the survey. In many types of research, such encouragement is not necessary either because participants do not know they are in a study (as in naturalistic observation) or because they are part of a subject pool and have already shown their willingness to participate by signing up and showing up for the study. Survey research usually catches respondents by surprise when they answer their phone, go to their mailbox, or check their e-mail—and the researcher must make a good case for why they should agree to participate. Thus, the introduction should briefly explain the purpose of the survey and its importance, provide information about the sponsor of the

survey (university-based surveys tend to generate higher response rates), acknowledge the importance of the respondent's participation, and describe any incentives for participating.

The second function of the introduction is to establish informed consent. Remember that this involves describing to respondents everything that might affect their decision to participate. This includes the topics covered by the survey, the amount of time it is likely to take, the respondent's option to withdraw at any time, confidentiality issues, and so on. Written consent forms are not always used in survey research (when the research is of minimal risk and completion of the survey instrument is often accepted by the IRB as evidence of consent to participate), so it is important that this part of the introduction be well documented and presented clearly and in its entirety to every respondent.

The introduction should be followed by the substantive questionnaire items. But first, it is important to present clear instructions for completing the questionnaire, including examples of how to use any unusual response scales. Remember that the introduction is the point at which respondents are usually most interested and least fatigued, so it is good practice to start with the most important items for purposes of the research and proceed to less important items. Items should also be grouped by topic or by type. For example, items using the same rating scale (e.g., a 5-point agreement scale) should be grouped together if possible to make things faster and easier for respondents. Demographic items are often presented last because they are least interesting to participants but also easy to answer in the event respondents have become tired or bored. Of course, any survey should end with an expression of appreciation to the respondent.

Conducting Surveys

Surveys are famously sensitive to the sampling methodology used to recruit participants into the research study. The sampling methods discussed in Chapter 11 provide the technical terms for a variety of approaches to this challenge. The main concern with being able to draw robust inferences from surveys arise from problems of **sampling bias**.

Probability sampling was developed in large part to address the issue of sampling bias. Sampling bias occurs when a sample is selected in such a way that it is not representative of the entire population and therefore produces inaccurate results. This bias was the reason that the Literary Digest straw poll was so far off in its prediction of the 1936 presidential election. The mailing lists used came largely from telephone directories and lists of registered automobile owners, which over-represented wealthier people, who were more likely to vote for Landon. Gallup was successful because he knew about this bias and found ways to sample less wealthy people as well.

There is one form of sampling bias that even careful random sampling is subject to. It is almost never the case that everyone selected for the sample actually responds to the survey. Some may have died or moved away, and others may decline to participate because they are too busy, are not interested in the survey topic, or do not participate in surveys on principle. If these survey non-responders differ from survey responders in systematic ways, then this difference can produce **non-response bias**. For example, in a mail survey on alcohol consumption, researcher Vivienne Lahaut and colleagues found that only about half the sample responded after the initial contact and two follow-up reminders (Lahaut, Jansen, van de Mheen, Garretsen, 2002).The danger here is that the half who responded might have different patterns of alcohol consumption than the half who did not, which could lead to inaccurate conclusions on the part of the researchers. So to test for non-response bias, the researchers later made unannounced visits to the homes of a subset of the non-responders—coming back up to five times if they did not find them at home. They found that the original non-responders

included an especially high proportion of abstainers (nondrinkers), which meant that their estimates of alcohol consumption based only on the original responders were too high.

Although there are methods for statistically correcting for non-response bias, they are based on assumptions about the non-responders—for example, that they are more similar to late responders than to early responders—which may not be correct. For this reason, the best approach to minimizing non-response bias is to minimize the number of non-responders—that is, to maximize the response rate. There is a large research literature on the factors that affect survey response rates (Groves et al., 2004). In general, in-person interviews have the highest response rates, followed by telephone surveys, and then mail and Internet surveys. Among the other factors that increase response rates are sending potential respondents a short pre-notification message informing them that they will be asked to participate in a survey in the near future and sending simple follow-up reminders to non-responders after a few weeks. The perceived length and complexity of the survey can also make a difference, which is why it is important to keep survey questionnaires as short, simple, and on topic as possible. Finally, offering an incentive—especially cash—is a reliable way to increase response rates. However, ethically, there are limits to offering incentives that may be so large as to be considered coercive.

The four main ways to conduct surveys are through in-person interviews, by telephone, through the mail, and over the internet. As with other aspects of survey design, the choice depends on both the researcher's goals and the budget. In-person interviews have the highest response rates and provide the closest personal contact with respondents. Personal contact can be important, for example, when the interviewer must see and make judgments about respondents, as is the case with some mental health interviews. But in-person interviewing is by far the most costly approach. Telephone surveys have lower response rates and still provide some personal contact with respondents. They can also be costly but are generally less so than in-person interviews. Traditionally, telephone directories have provided fairly

comprehensive sampling frames. However, this trend is less true today as more people choose to only have cell phones and do not install land lines that would be included in telephone directories. Mail surveys are less costly still but generally have even lower response rates—making them most susceptible to non-response bias.

Not surprisingly, internet surveys are becoming more common. They are increasingly easy to construct and use. Although initial contact can be made by mail with a link provided to the survey, this approach does not necessarily produce higher response rates than an ordinary mail survey. A better approach is to make initial contact by email with a link directly to the survey. This approach can work well when the population consists of the members of an organization who have known email addresses and regularly use them (e.g., a university community). For other populations, it can be difficult or impossible to find a comprehensive list of email addresses to serve as a sampling frame. Alternatively, a request to participate in the survey with a link to it can be posted on websites known to be visited by members of the population. But again it is very difficult to get anything approaching a random sample this way because the members of the population who visit the websites are likely to be different from the population as a whole. However, internet survey methods are in rapid development. Because of their low cost, and because more people are online than ever before, internet surveys are likely to become the dominant approach to survey data collection in the near future.

Surveys as Research Instruments

There are a wide variety of established questionnaires that have been developed and extensively tested and used across studies to establish a reliable and valid measure of a specific set of underlying constructs. A well-established survey measure with a strong history of use in research will often be referred to as a research instrument. These will be published as a list of the questions used in the measure and a scoring system for combining

responses into a single quantitative score (as in the Self-Esteem measure described in Chapter 2). There are a large number of these established measures you may encountered in reading the research literature.

Some of these existing measures, particularly those that have applications in clinical psychology, are proprietary. This means that a publisher owns the rights to them and that you would have to purchase them. These include many standard intelligence tests, the Beck Depression Inventory, and the Minnesota Multiphasic Personality Inventory (MMPI). Details about many of these measures and how to obtain them can be found in other reference books, including Tests in Print and the Mental Measurements Yearbook. There are also tools implemented in assessment technology like tablet computers that provide reliable, consistent assessments like the NIH ToolBox of neurobehavioral assessments.

In planning research, it is generally a good idea to use an existing measure that has been used successfully in previous research instead of attempting to develop your own. Among the advantages are that (a) you save the time and trouble of creating your own, (b) there is already some evidence that the measure is valid (if it has been used successfully), and (c) your results can more easily be compared with and combined with previous results. In fact, if there already exists a reliable and valid measure of a construct, other researchers might expect you to use it unless you have a good and clearly stated reason for not doing so.

If you choose to use an existing measure, you may still have to choose among several alternatives. You might choose the most common one, the one with the best evidence of reliability and validity, the one that best measures a particular aspect of a construct that you are interested in or even the one that would be easiest to use. For example, the Ten-Item Personality Inventory (TIPI) is a self-report questionnaire that measures all the Big Five personality dimensions with just 10 items (Gosling, Rentfrow & Swann, 2003). It is not as reliable or valid as longer and more comprehensive measures, but a researcher might choose to use it when testing time is severely limited.

Creating a New Measure

Instead of using an existing measure, you might want to create your own. Perhaps there is no existing measure of the construct you are interested in or existing ones are too difficult or time-consuming to use. Or perhaps you want to use a new measure specifically to see whether it works in the same way as existing measures—that is, to evaluate convergent validity. In this section, we consider some general issues in creating new measures that apply equally to self-report, behavioral, and physiological measures.

First, be aware that most new measures in psychology are really variations of existing measures, so you should still look to the research literature for ideas. Perhaps you can modify an existing questionnaire, create a paper-and-pencil version of a measure that is normally computerized (or vice versa), or adapt a measure that has traditionally been used for another purpose. For example, the famous Stroop task (Stroop, 1935)—in which people quickly name the colors that various color words are printed in—has been adapted for the study of social anxiety. People high in social anxiety are slower at color naming when the words have negative social connotations such as "stupid" (Amir, Freshman, & Foa, 2002).

When you create a new measure, you should strive for simplicity, aiming to keep the measure brief to avoid boring or frustrating your participants to the point that their responses start to become less reliable and valid. The need for brevity, however, needs to be weighed against the fact that it is nearly always better for a measure to include multiple items rather than a single item. There are two reasons for this. One is a matter of content validity. Multiple items are often required to cover a construct adequately. The other is a matter of reliability. People's responses to single items can be influenced by all sorts of irrelevant factors—misunderstanding the particular item, a momentary distraction, or a simple error such as checking the wrong response option. But when several responses are summed or averaged, the effects of these irrelevant factors tend to cancel each other out to produce more reliable scores. When using multiple items this way, there will typically be a way to

combine them into a single overall score by summing or averaging.

The method of scoring the items is often part of the procedure for implementing the new survey questions. Much of this will use ideas from previous discussions of experimental control, avoiding bias in responses and considering demand characteristics on participants. Although informed consent requires telling participants what they will be doing, it does not require revealing your hypothesis or other information that might suggest to participants how you expect them to respond. A questionnaire designed to measure financial responsibility need not be titled "Are You Financially Responsible?" It could be titled "Money Questionnaire" or have no title at all. Finally, the effects of your expectations can be minimized by arranging to have the measure administered by a helper who is "blind" or unaware of its intent or of any hypothesis being tested. Regardless of whether this is possible, you should standardize all interactions between researchers and participants—for example, by always reading the same set of instructions word for word.

When using questionnaires that ask about sensitive or personal questions, methods can be used to guarantee participants' anonymity and make clear to them that you are doing so. If you are testing them in groups, be sure that they are seated far enough apart that they cannot see each other's responses. You can even allow them to seal completed questionnaires into individual anonymous envelopes or put them into a drop box where they immediately become mixed with others' questionnaires.

Evaluating a Measure

Every new measure needs to be thoroughly evaluated in terms of its **reliability** and **validity**. These terms are used here in the same spirit as when they are applied to research design. A reliable measure should produce similar results when used multiple times. A valid measure is thought to be an effective assessment of the intended construct.

Instrument reliability requires that it be consistent when used to measure a construct. Psychologists consider three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability).

Test-retest reliability is the extent to which the measure is consistent across time. For example, intelligence is generally thought to be consistent across time. A person who is highly intelligent today will be highly intelligent next week. This means that any good measure of intelligence should produce roughly the same scores for this individual next week as it does today. Clearly, a measure that produces highly inconsistent scores over time cannot be a very good measure of a construct that is supposed to be consistent.

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the same group of people at a later time, and then looking at the test-retest correlation between the two sets of scores. This is typically done by computing the correlation coefficient between tests (Chapter 16).

Again, high test-retest correlations make sense when the construct being measured is assumed to be consistent over time, which is the case for intelligence, self-esteem, and the Big Five personality dimensions. But other constructs are not assumed to be stable over time. The very nature of mood, for example, is that it changes. A measure of mood that produced a low test-retest correlation over a period of a month would not be a cause for concern.

Another kind of reliability is **internal consistency**, which is the consistency of people's responses across the items on a multiple-item measure. In general, all the items on such measures are supposed to reflect the same underlying construct, so people's scores on those items should be correlated with each other. On the Rosenberg Self-Esteem Scale, people who agree that they are a person of worth should tend to agree that they have a number of good qualities. If people's responses to the different items are not correlated with each other, then it would no longer make sense to claim that they are all measuring the same underlying construct. This is as true for behavioral

and physiological measures as for self-report measures. For example, people might make a series of bets in a simulated game of roulette as a measure of their level of risk seeking. This measure would be internally consistent to the extent that individual participants' bets were consistently high or low across trials.

Like test-retest reliability, internal consistency can only be assessed by collecting and analyzing data. One approach is to look at a split-half correlation. This involves splitting the items into two sets, such as the first and second halves of the items or the even- and odd-numbered items. Then a score is computed for each set of items, and the relationship between the two sets of scores is examined. A more elaborate version of this is to use a statistic called Cronbach's α (the Greek letter alpha). Conceptually, α is the mean of all possible split-half correlations for a set of items. For example, there are 252 ways to split a set of 10 items into two sets of five. Cronbach's α would be the mean of the 252 split-half correlations. Note that this is not how α is actually computed, but it is a correct way of interpreting the meaning of this statistic. Again, a value of +.80 or greater is generally taken to indicate good internal consistency.

Many behavioral measures involve significant judgment on the part of an observer or a rater. **Inter-rater reliability** is the extent to which different observers are consistent in their judgments. For example, if you were interested in measuring university students' social skills, you could make video recordings of them as they interacted with another student whom they are meeting for the first time. Then you could have two or more observers watch the videos and rate each student's level of social skills. To the extent that each participant does, in fact, have some level of social skills that can be detected by an attentive observer, different observers' ratings should be highly correlated with each other. Inter-rater reliability would also have been measured in Bandura's Bobo doll study. In this case, the observers' ratings of how many acts of aggression a particular child committed while playing with the Bobo doll should have been highly positively correlated. Interrater reliability is often assessed using Cronbach's α when the judgments

are quantitative or an analogous statistic called Cohen's κ (the Greek letter kappa) when they are categorical.

Instrument Validity

Validity is the extent to which the scores from a measure represent the construct they are intended to. But how do researchers make this judgment? We have already considered one factor that they take into account—reliability. When a measure has good test-retest reliability and internal consistency, researchers should be more confident that the scores represent what they are supposed to. There has to be more to it, however, because a measure can be extremely reliable but have no validity whatsoever. As an absurd example, imagine someone who believes that people's index finger length reflects their self-esteem and therefore tries to measure self-esteem by holding a ruler up to people's index fingers. Although this measure would have extremely good test-retest reliability, it would have absolutely no validity. The fact that one person's index finger is a centimeter longer than another's would indicate nothing about which one had higher self-esteem.

Discussions of validity usually divide it into several distinct "types." But a good way to interpret these types is that they are other kinds of evidence—in addition to reliability—that should be taken into account when judging the validity of a measure. Here we consider three basic kinds: face validity, content validity, and criterion validity.

Face validity is the extent to which a measurement method appears "on its face" to measure the construct of interest. Most people would expect a self-esteem questionnaire to include items about whether they see themselves as a person of worth and whether they think they have good qualities. So a questionnaire that included these kinds of items would have good face validity. The finger-length method of measuring self-esteem, on the other hand, seems to have nothing to do with self-esteem and therefore has poor face validity. Although face validity can be assessed quantitatively—for example, by having a large sample of people rate a measure in terms of whether it appears to measure what it is intended to—it is usually assessed

informally.

Face validity is at best a very weak kind of evidence that a measurement method is measuring what it is supposed to. One reason is that it is based on people's intuitions about human behavior, which are frequently wrong. It is also the case that many established measures in psychology work quite well despite lacking face validity. The Minnesota Multiphasic Personality Inventory-2 (MMPI-2) measures many personality characteristics and disorders by having people decide whether each of over 567 different statements applies to them—where many of the statements do not have any obvious relationship to the construct that they measure. For example, the items "I enjoy detective or mystery stories" and "The sight of blood doesn't frighten me or make me sick" both measure the suppression of aggression. In this case, it is not the participants' literal answers to these questions that are of interest, but rather whether the pattern of the participants' responses to a series of questions matches those of individuals who tend to suppress their aggression.

Content validity is the extent to which a measure "covers" the construct of interest. For example, if a researcher conceptually defines test anxiety as involving both sympathetic nervous system activation (leading to nervous feelings) and negative thoughts, then his measure of test anxiety should include items about both nervous feelings and negative thoughts. Or consider that attitudes are usually defined as involving thoughts, feelings, and actions toward something. By this conceptual definition, a person has a positive attitude toward exercise to the extent that they think positive thoughts about exercising, feels good about exercising, and actually exercises. So to have good content validity, a measure of people's attitudes toward exercise would have to reflect all three of these aspects. Like face validity, content validity is not usually assessed quantitatively. Instead, it is assessed by carefully checking the measurement method against the conceptual definition of the construct.

Criterion validity is the extent to which people's scores on a measure are correlated with other variables (known as criteria) that one would expect

them to be correlated with. For example, people's scores on a new measure of test anxiety should be negatively correlated with their performance on an important school exam. If it were found that people's scores were in fact negatively correlated with their exam performance, then this would be a piece of evidence that these scores really represent people's test anxiety. But if it were found that people scored equally well on the exam regardless of their test anxiety scores, then this would cast doubt on the validity of the measure.

A criterion can be any variable that one has reason to think should be correlated with the construct being measured, and there will usually be many of them. For example, one would expect test anxiety scores to be negatively correlated with exam performance and course grades and positively correlated with general anxiety and with blood pressure during an exam. Or imagine that a researcher develops a new measure of physical risk taking. People's scores on this measure should be correlated with their participation in "extreme" activities such as snowboarding and rock climbing, the number of speeding tickets they have received, and even the number of broken bones they have had over the years. When the criterion is measured at the same time as the construct, criterion validity is referred to as concurrent validity; however, when the criterion is measured at some point in the future (after the construct has been measured), it is referred to as predictive validity (because scores on the measure have "predicted" a future outcome).

Criteria can also include other measures of the same construct. For example, one would expect new measures of test anxiety or physical risk taking to be positively correlated with existing established measures of the same constructs. This is known as convergent validity. The use of convergent validity is obviously challenging for cases where a complete new measure is being developed. In some cases, the approach is used of showing that the measure does not capture a construct measured by a different instrument.

Discriminant validity, on the other hand, is the extent to which scores on a measure are not correlated with measures of variables that are conceptually distinct. For example, self-esteem is a general attitude toward the self that is fairly stable over time. It is not the same as mood, which is how good or bad

one happens to be feeling right now. So people's scores on a new measure of self-esteem should not be very highly correlated with their moods. If the new measure of self-esteem were highly correlated with a measure of mood, it could be argued that the new measure is not really measuring self-esteem; it is measuring mood instead.

When they created the Need for Cognition Scale, Cacioppo and Petty also provided evidence of discriminant validity by showing that people's scores were not correlated with certain other variables. For example, they found only a weak correlation between people's need for cognition and a measure of their cognitive style—the extent to which they tend to think analytically by breaking ideas into smaller parts or holistically in terms of "the big picture." They also found no correlation between people's need for cognition and measures of their test anxiety and their tendency to respond in socially desirable ways. All these low correlations provide evidence that the measure is reflecting a conceptually distinct construct.

It should be clear that establishing a new measure that has high reliability and robust validity is typically the product of an extensive program of psychological research. The recommendation here to generally prefer using existing, published measures is to take advantage of the work that previous researchers have already invested in doing this. At the same time, simple rating scales presented via surveys are often highly effective methods for collecting information about participants attitudes, opinions or intentions. These less formal approaches typically have very high face validity but the lack of prior research means that it may be necessary to consider limitations of the research of this kind of measure together with the standard concerns about sampling and possible bias from non-responders.

Key Takeaways

- Survey research features the use of self-report measures on carefully selected samples. It is a flexible approach that can be used to study a wide variety of basic and applied research questions.
- Survey research has its roots in applied social research, market research, and election polling. It has since become an important approach in many academic disciplines, including political science, sociology, public health, and, of course, psychology.
- Survey research involves asking respondents to self-report on their own thoughts, feelings, and behaviors.
- Most survey research is non-experimental in nature (it is used to describe variables or measure statistical relationships between variables) but surveys can also be used to measure dependent variables in true experiments.
- Responding to a survey item is itself a complex cognitive process that involves interpreting the question, retrieving information, making a tentative judgment, putting that judgment into the required response format, and editing the response.
- Survey responses are subject to numerous context effects due to question wording, item order, response options, and other factors. Researchers should be sensitive to such effects when constructing surveys and interpreting survey results.
- Survey items are either open-ended or closed-ended. Open-ended items simply ask a question and allow respondents to answer in whatever way they want. Closed-ended items ask a question and provide several response options that respondents must choose from.
- Use verbal labels instead of numerical labels although the responses can be converted to numerical data in the analyses.
- According to the BRUSO model, questionnaire items should be brief, relevant, unambiguous, specific, and objective.
- Survey research usually involves probability sampling, in which each member of

the population has a known probability of being selected for the sample. Types of probability sampling include simple random sampling, stratified random sampling, and cluster sampling.

- Sampling bias occurs when a sample is selected in such a way that it is not representative of the population and therefore produces inaccurate results. The most pervasive form of sampling bias is non-response bias, which occurs when people who do not respond to the survey differ in important ways from people who do respond. The best way to minimize non-response bias is to maximize the response rate by prenotifying respondents, sending them reminders, constructing questionnaires that are short and easy to complete, and offering incentives.
- Surveys can be conducted in person, by telephone, through the mail, and on the internet. In-person interviewing has the highest response rates but is the most expensive. Mail and internet surveys are less expensive but have much lower response rates. Internet surveys are likely to become the dominant approach because of their low cost.

Exercises

- Discussion: Think of a question that each of the following professionals might try to answer using survey research.
    - a social psychologist
    - an educational researcher
    - a market researcher who works for a supermarket chain
    - the mayor of a large city
    - the head of a university police force
- Discussion: Write a survey item and then write a short description of how someone might respond to that item based on the cognitive model of survey responding (or choose any item on the Rosenberg Self-Esteem Scale at http://www.bsos.umd.edu/socy/research/rosenberg.htm).
- Practice: Write survey items for each of the following general questions. In some cases, a series of items, rather than a single item, might be necessary.
    - How much does the respondent use Facebook?
    - How much exercise does the respondent get?
    - How likely does the respondent think it is that the incumbent will be re-elected in the next presidential election?
    - To what extent does the respondent experience "road rage"?
- Discussion: If possible, identify an appropriate sampling frame for each of the following populations. If there is no appropriate sampling frame, explain why.
    - students at a particular university
    - adults living in the state of Washington
    - households in Pullman, Washington
    - people with low self-esteem