# Designing Surveys for Evaluations and Research

Randall S. Davies

# Table of Contents


## Randall S. Davies

Brigham Young University


Dr. Randall S. Davies is an associate professor of Instructional Psychology & Technology at Brigham Young University (BYU), where he teaches courses in educational inquiry, measurement, and evaluation. He previously worked at Indiana University South Bend as an Assistant Professor of Educational Research and at IU Bloomington in the Center for Evaluation and Education Policy. He received his Ph.D. in Instructional Psychology & Technology from BYU as well as a B.Ed. in secondary education and a B.Sc. in mathematics & computing science from the University of Alberta.

# An Overview of Survey Use

## Survey Basics

Surveys are often used for evaluations and research to collect data. The concept of creating and administering a survey is deceptively simple. You probably have been asked to take several surveys, so you (and many others) most likely feel you understand the idea well enough. Taking a survey means answering questions, and administering a survey involves asking questions. As simple as this process appears, you probably won't be surprised to find that there is a bit more to it than that.

Many of the definitions provided for the term *survey* involve describing the function or purpose of the process. In the social sciences, "to survey" means getting a general picture or description of some identifiable group of individuals. However, a survey also refers to a data collection instrument used to gather information from the group of individuals you want to understand better. The term *survey* is used to describe the activity but also the data collection instrument. *Survey Research* refers to any study or evaluation that uses a survey as its primary data collection method.

When people talk about using a survey, they usually mean a sample survey. In a *sample survey*, researchers select only a portion of some target population with the intention of generalizing the results to all those in that population. When researchers survey the entire population, they are conducting a *census*. However, in many cases, surveying the entire population is not feasible; as an alternative, researchers collect information from some individuals to gain insights into the entire group. If done correctly, surveying a portion of the population will produce a reasonable description of the entire population.

### Definitions

A **survey** is a self-report data collection instrument designed to get information from individuals in a specific group.

A **sample survey** refers to any survey that gathers information from a sample group in order to make generalizations about the group's population.

A **representative sample** is one that provides a reasonable representation of the population targeted in the study.

**Survey research** refers to any study or evaluation that uses a survey as its primary data collection method.

Several aspects of a survey research study determine its quality, but the overall quality of the study will primarily depend on the quality of the survey instrument used to collect information. More specifically, the quality of the items on the survey instruments has a significant impact on the overall quality of the study. It is relatively easy to write survey

questions, but it takes effort and skill to craft good ones. It is important to make sure the survey asks all the right questions, in the right way, and is administered to the right people.

The results from a survey may be used to gain insights about a specific group of individuals, but they also can be used to make generalizations about the population the sample comes from. The degree to which the generalizations are valid depends on the sampling methods used and the response rate. In order for the results to be generalizable, the sample must be a *representative sample*, meaning those responding to the survey are a reasonable representation of the population targeted in the study.

Survey research is conducted for a variety of purposes. Often the researcher wants only to better understand the characteristics of individuals in a specific population. Characteristics of interest might include attitudes, opinions, trends, or perceptions. The only way to get these data is to ask individuals directly. This means the respondents must be willing and able to answer the questions truthfully and accurately; the value of any survey research study depends on this. If you can get a more accurate measure of a characteristic in another way, a self-report survey may not be your best choice for data collection.

Understandably, critics of survey research point out these limitations as a reason to question the results of many studies. But with any assessment tool, there will be measurement error. Proponents of survey research point out that having a reasonable estimate, even if not perfect, is often better than having no data at all. However, when conducting survey research you must recognize, attend to, and address threats to validity by designing and creating good surveys.

# Types of Surveys

Surveys can take various forms, but the most common form is a questionnaire. In this course, we will also discuss how to create a scale. The basic difference between these two types of surveys is the relationship between the items on the survey instrument.

Each item stands on its own in a questionnaire and is reported separately. Each item provides data that can help answer different parts of the research question. In a scale, the items work together to describe a specific construct or an attribute of an individual participant. Each item in the scale focuses on one aspect of the construct. In a scale, the item responses are combined to create a score. The construct being measured is often affective in nature in that it describes some attribute of the individual respondent. For example, a scale might measure typical feelings, attitudes, or perceptions about something. The resulting score is an indication of the magnitude and direction of the attribute being measured.

With both survey forms, summarizing the results of those individuals participating in the survey can be used to describe those in the population (i.e., the proportion of individuals within a specific category or, in the case of a scale, the average score of individuals within the population). However, a scale might also be used as a measurement instrument designed to describe some characteristic of an individual but not necessarily a description of the targeted population.

## Chapter Summary

- A **survey** is a self-report data collection instrument.
- Surveys typically take the form of a questionnaire but can also be designed as a scale (or a mixture of the two).
- In a questionnaire, each item acts independently of the others and is analyzed and reported separately.
- In a scale, several items work together to provide a score that represents the degree to which an individual possesses some characteristic.
- Surveys are often administered online (for convenience and cost reasons) but they can be administered in other ways, for example, as an interview.
- **Survey research** includes any study that uses a survey as its main data collection tool.
- Although a survey may not be the ideal method for studying every research question, a well-designed survey can enhance our understanding of just about any issue.
- A survey is often the only means available to get the information needed to answer certain research questions.
- Most survey research is non-experimental (i.e., descriptive research) but surveys are also used in experimental and quasi-experimental research. Survey research is mostly seen as a quantitative research method.
- Surveying the entire population is called a census.
- Sample surveys obtain information from a representative group in the target population then attempt to make generalizable conclusions about those in the population.
- In order to make valid generalizations, the sample used must be a representative sample of the target population.
- Conducting high-quality survey research will depend on asking the right questions, in the right way, to the right people.
- It is easy to create a flawed survey. Without careful attention to overall survey design and the creation of the survey items, the study will likely produce invalid results.

## Discussion Questions

1. Under what condition would a survey be a good choice for your research?
2. When would a scale be needed rather than a questionnaire?

# The Design Process

## Overview and Importance

The goal for any research is to get accurate and comprehensive information that can be used to answer the research questions. Survey research is no different. However, creating a survey that provides useful results takes time and effort. Unfortunately, many researchers fail to take adequate time to validate their survey instruments. One of the worst things for researchers is to get their survey data back and realize they should have created the survey differently. Without valid data, the study is ruined. And while it is unlikely that any survey you might create will be perfect, following a rigorous survey design process will increase the likelihood that the data you obtain will be useful.

### Instrument Validation

Unfortunately, the survey design process is not understood very well; many people do not know how to accomplish the task properly, and too often the process is not carried out thoroughly. More importantly, many do not understand that proper survey design involves instrument validation.

The term validity has a slightly different meaning when applied to assessment (i.e., testing) and surveys (i.e., self-report data collection). Validity in assessment generally refers to the degree to which a test measures what it was designed to measure. This definition can apply to surveys as well, however for survey research, validity also refers to the quality, completeness, and usefulness of the data. More specifically, validating a survey instrument is a rigorous process intended to verify that the data obtained from administering a survey will answer the research questions adequately. This does not mean the information obtained from the administration of a validated survey will be perfect. There are many ways in which the information obtained can be biased regardless of the quality of the survey (see Survey Error and Response Bias). Validation simply decreases the potential for response bias and helps us avoid mistakes that can result in survey error. Again, the validation process does not guarantee perfect data; rather, once a thorough validation process is completed, the researcher can be confident that the instrument has the potential to produce useful information. The validation process attempts to make sure the survey can be used to accurately gather all the data it was designed to obtain.

### Phases in the Survey Design Process

The survey design (i.e., validation) process involves several steps or phases. They are outlined below. Subsequent chapters then deal with specific aspects of each phase in greater detail.

**Conceptualization Phase.** This is a theoretical endeavor that affects all the other steps in the process. The aim of this phase is to clearly identify the study's purpose and research questions. Having a clear understanding of the purpose and research questions helps us figure out what data is needed.

**Administration Planning.** Although implementation is not technically a part of the design process, this phase involves planning the survey administration to ensure adequate sampling and sufficient response rates. Understanding how the survey will be administered will also affect item creation.

**Data Analysis and Report Planning.** This is a planning step that will guide how data is collected. Understanding how the data will be analyzed and reported improves the likelihood that you get all the data you need and that the information obtained will be adequate for the statistical analysis needed to answer the research questions.

**Item Creation and Response Scale Design.** This phase involves development. Both survey items and response scales are created. The goal is to make items that will produce the information required.

**Refinement and Pilot Testing.** The objective of this phase is to make sure mistakes have not been made before administering the final survey. Based on the results, changes to specific items can be made. Depending on the results, several iterations of the phase may be needed.

### Chapter Summary

- There are no perfect survey instruments, but the likelihood that a survey will not be flawed is increased if a rigorous design process is followed.
- A proper survey design process will include instrument validation.
- Survey validation involves several steps or phases intended to verify that the data obtained from administering a survey will answer the research questions adequately.
- The validation process does not guarantee perfect data. However, once a thorough validation process is completed, the researcher can be confident that the instrument has the potential to produce useful information.
- The design phases include conceptualizing the purpose and establishing research questions. It also requires planning the administration, data analysis, and reporting. Instrument creation should also include a refinement and testing cycle.

### Discussion Questions

1. How does failing to complete each of the steps in the validation process affect the results of a survey?
2. Pick one of the validation steps and explain how skipping or not completing the step may affect the quality of the survey.

## Section Contents

# Conceptualization Phase

## Basic Processes and Principles

Survey Research

Prior to creating survey items, you need to determine the precise topic and purpose for the study as well as the research questions you wish to answer. Careful consideration of the overall study's purpose and questions will help identify whether a survey is the best way to collect data, what data are needed, and who should take the survey.

## Research Topics

A good study starts by clearly and succinctly identifying the specific topic that will be addressed. A good research topic will be interesting, important, and researchable. The topic may be of interest simply because you are an expert in the field and this is your chosen field of study. However, a topic may also be of interest to you and others because current events have identified the topic as something we need to better understand. Another characteristic of a good research topic is that of importance. Importance can be a somewhat subjective factor; however, a research topic has significance if it has the potential to add to our general understanding of theory and/or improve practice. The issue of whether a research topic is researchable has to do with the scope of the topic and the degree to which the researcher can manage the study. Narrowing a topic makes it more manageable, which then makes the topic researchable. In addition, making sure the researcher can gain access to or gather relevant data will also determine whether the topic can be researched properly.

There are many ways to find a good topic. However, a review of past research is generally a good idea. Conducting a review of published research will identify potential areas for new research or the need to replicate previous findings. For example, studies may have explored a specific topic in the past but, over time and given changing circumstances, the research may need to be replicated to determine what (if anything) has changed or to verify the results. A good topic might also come from personal experience. A researcher may find that they need to better understand some aspect of their work in order to improve their practice in a certain area.

A research topic might address attitudes, perceptions, or some personal characteristic of individuals within a specific population. The topic to be explored might involve some social condition or circumstance. A well-written topic statement will identify the variables the research plans to study, possibly including the hypothesized relationship between variables. The topic statement should also delineate the scope of the research by describing important characteristics of the participants. Depending on the complexity and familiarity of the topic, precisely stating the topic in a single sentence may be a challenge. Remember, specific terms and definitions are not always understood universally; you may need to take additional time to precisely define the topic, both theoretically and operationally, as part of the introduction or through a more in-depth review of previous research.

### Topic Statement Example

Consider the following topic statement. “This study investigated counseling services available to first-year students.” In this example the general topic involves counseling services, specifically the availability of such services. The participant population for the study is narrowed to include first-year students. However, it is unclear what is meant by first-year students and what type of counseling services will be considered. A better version of the statement would specifically state the type of counseling services (e.g., academic, emotional, or psychological). It would also more clearly define the scope in terms of which students would be involved (e.g., first grade elementary, high school, or university students). A better version of this topic statement might be, “This study investigated counseling services for first-year university students who suffer from depression and anxiety.” Should the researcher desire, the scope might be more narrowly defined by specifying a specific university, type of university, program within the university, or subgroup of the student population. It could be expanded to include students other than first-year students.

# Research Purpose

Once the specific topic has been defined, a purpose for the study must be established. The purpose of the study is different from the purpose of the survey. Simply put, the purpose of a survey is to gather information. The study’s purpose will be to address a specific problem or resolve a particular issue. The purpose of the study is closely associated with the research questions in that the answers to the research questions (obtained from the survey) will help solve the problem or address the issues.

A clear purpose statement will help focus the survey so the survey will concentrate on what is needed rather than tangential and irrelevant data collection which might distract from or sidetrack the actual purpose of the study. The topic statement might include the purpose; in other words, the statement might indicate the problem the research attempts to solve. However, depending on the situation, a separate purpose statement is most often provided.

A complete purpose statement should include an explanation as to why the problem is important (i.e., providing a justification without simply stating the research is important) and a description of how a better understanding of the topic might contribute to generalizable theory or specific practice (i.e., the potential benefit for conducting the study). Basically, you need to determine what problem is being addressed and how the research you plan to conduct will help solve the problem.

The purpose for a research study might be to

- describe the prevalence of a specific condition within a targeted population,
- investigate the relationship between two variables,
- establish whether (i.e., how and to what degree) a specific intervention or program is beneficial, or
- understand which factors (i.e., independent variables) influence a specific dependent variable (e.g., behavior, condition, or outcome).

# Proving vs. Providing Evidence

Rarely will anyone use the word “prove” when describing the purpose of a study. Even when attempting to establish a potential causal link or attribute some outcome to a specific intervention, the general purpose of any inquiry is to gather evidence that will help those investigating answer the questions posed in a study. The evidence obtained will either support or refute hypothesized conjectures, never prove them.

### Research Purpose Example

You will recall that in the counseling services topic statement example, the precise purpose for the study is not fully specified. In this example, the purpose could have been simply to determine students' awareness of the services available to them; however, a more interesting purpose might have included the need for or adequacy of such services. A potential topic and purpose statement for the counseling services study might have been stated as follows.

> *This study investigated counseling services available to first-year university students who suffer from depression and anxiety. Given the increased number of students who report suffering from stress in their transition from high school to university, the purpose of this study was to determine students' awareness of counseling services, as well as the adequacy and accessibility of those services intended to alleviate excessive anxiety and depression. A better understanding of the current situation will be used to set university policy and address any issues that need to be resolved.*

# Research Questions

Good research questions align directly with the purpose of the study. Answering the research questions helps resolve the problem being addressed by the study. A good research question allows for rich, in-depth answers. Multiple questions could be poised, but the questions should address the overall purpose of the study, not specific survey questions. Listing too many research questions can be a problem when reporting because reporting survey research will already be a lengthy endeavor.

## Impossible Questions

There are some research questions that should be avoided. Questions that involve values, morals, or ethics can be a challenge—especially when attempting to determine which value-based moral or ethical stances everyone *should* take. These questions cannot be answered through empirical research because, like opinions and feelings, there is no universally accepted correct way to think or feel. Differences in the way we think or feel are an essential part of being human and having agency. This does not mean that topics involving values, morals, and ethics cannot be studied. Research questions might address the prevalence of specific value-based moral or ethical positions being held by individuals in distinct societal groups. One might ask about the rationale or reasons for why people hold such positions. One might even explore the outcomes that result from holding specific value-laden positions, both intended and unintended. Answering questions in this unbiased way can provide information that has the potential to enhance dialogue regarding important social, emotional, and religious issues. However, when choosing the research questions for a study, one should avoid asking which value-based moral or ethical stance people *should* take or opinion people *should* espouse. These are impossible questions.

## Research Questions Example

Continuing with the counseling services example, a researcher could ask “Do participants value the counseling services provided?” However, the answer to this question is a yes/no response (i.e., a binary or dichotomous response). Values are rarely dichotomous in nature, rather they fall on a continuum. A better question might be “To what extend and in what ways do participants value the counseling service programs?” The researcher will also need to decide whether values should be the intended focus. Rather than use the term "value,” researchers might consider determining participants’ perceptions of effectiveness. Participants may value having access to counseling services but not feel they are particularly effective or accessible. Depending on the purpose for the study, the study may attempt to answer multiple research questions. A second research question might involve gaining a better understanding of which services are of greatest need. However, it is generally good practice to keep the research question at a broader, more general level rather than the specific—unless it is determined that doing so would be needed in order to narrow the scope of the study or directly resolve the issue being addressed. For example, providing research questions that targeted a lengthy list of specific counseling services may not be advisable; a more general research question would suffice. However, if specific services are truly the focus of the study, it could and should be specified in the purpose statement and methods section of any planned research.

Possible research questions for the counseling services study example might include.

- To what degree are students aware of the counseling services currently being offered?
- What additional services do students feel they need?
- How willing are students to avail themselves of specific counseling services?
- To what degree do students feel needed counseling services are accessible to them?
- To what degree do students feel the available services adequately meet their needs?

## Chapter Summary

- Before attempting to write survey items, the topic, purpose, and research questions should be clearly defined and stated.
- A good research topic will be interesting, important, and researchable.
- A study’s purpose will address a specific problem or issue.
- It is never a good idea to presume the purpose of research is to prove anything. Research can only provide evidence that either supports or refutes a hypothesis.
- Good research questions align directly with the purpose of the study.
- Research questions involving morals, values, and ethics that are unlikely to have a universally accepted correct answer should be avoided. This does not mean these topics should not be studied, rather it is recommended that any impossible "should" questions be reframed in ways that they can provide information that has the potential to enhance dialogue and understanding.

## Discussion Questions

1. Why is it important to clearly define the topic, purpose, and research questions prior to writing items for the survey?
2. How are the research purpose and research questions related?

## Practice Tasks

1. Consider the following topic statement: “In this study we will explore the relationship between high school students’ feelings of safety and their scholastic effort.”

   What is the topic? How does the statement narrow and delineate the scope? What specific variables will be addressed? What aspect of the variables will be explored? How could it be improved?

2. Choose a potential topic. Write a sentence that clearly and succinctly states the topic. Evaluate the topic using the criteria of interest, importance, and research viability. Revise as needed.

3. For the topic you choose in Practice Task 2, create a purpose statement and potential research questions. Evaluate their potential adequacy, importance, and completeness. Revise as needed.

# Survey Administration Planning

## Basic Concepts and Principles

After the topic, purpose, and specific research questions have been determined, a few additional aspects of the research need to be planned. The first thing the researcher needs to determine is what data is needed. This is important so you can ask the right questions. The researcher must then decide whether a survey is the best way to collect the required data. You will get better data if you get data in the right way. If a survey is required, a determination needs to be made as to who should be invited to take the survey. Before you start creating survey items, you need to make sure you are asking the right people (i.e., those who can provide the information needed). As a general rule, you need to ask the right questions, to the right people, in the right way. This phase of the planning addresses these aspects of the research. Good planning will attempt to anticipate problems and attempt to prevent them.

## What Data Needs to Be Gathered?

Before you can decide whether a survey is a good choice for data collection, you have to determine what information is needed. This decision will also be framed by the purpose of the study. If done carefully, considering the purpose of the study will help you ensure that you get all the data you need to answer the research questions.

During this phase you will want to review existing literature on the topic addressed in the study. By doing this you will get a better idea regarding the data other researchers have gathered and perhaps data that was not but should have been obtained. Not only will this help you gain a better understanding of the topic and issues, but it will also help you understand what instruments others may have already created that can be used during the designing or implementing of your survey. This is especially important when designing a scale. For a scale, the theoretical definition of the construct being measured helps to establish an operational definition which is needed for identifying the component parts or aspects of the construct the survey will attempt to measure.

Considering how the data will be analyzed may also need to be part of the planning. It is one thing to report results descriptively for the entire participant group, but it is another to report differences between subgroups. For example, if you want to disaggregate the data for purposes of comparing responses from different groups, you need to make sure you have sufficient information that will allow you to homogeneously categorize individuals into distinct groups in a mutually exclusive manner. If the purpose of the study is to explore the relationship between two variables, then an accurate measure of the variables in question is needed. This may require a survey, or it may require some other type of measurement instrument.

### Data Identification Example

In the counseling services study example from the previous chapter, suppose you wanted to address the issue of need. This issue has various aspects, so you might want to determine the extent of the problem. However, perceptions of need can be personal or general. When a participant is asked about their needs and their perceptions of others' needs, you may get different answers (e.g., "I'm not depressed, but I know a lot of people who are."). A different approach might attempt to better understand whether students feel the available services adequately meet their personal needs and, more generally, others' needs. Information from both these queries (i.e., perceived need and availability) might be required if the study's purpose was to identify the relationship between personal and general perception (e.g., "Does an individual's personal situation influence perceived need?" or "Is the relationship different if the need is a personal or a general perception?"). Still another aspect of the issue might explore the perceived need for services that currently are not being provided. In addition, if someone was under considerable stress, doing poorly in school, or in a particularly challenging program, you might hypothesize that their feelings about counseling service adequacy may differ. If this were the case, you will need to obtain specific demographic and situational information. In any case, you need to carefully determine what information you will need to gather.

## Is a Survey Needed?

Deciding whether a survey is needed depends on whether self-report is the best way to obtain the information you need. A self-report survey is needed in situations where asking an individual is the only (or best) way to get accurate data. Sometimes the decision to use a survey is clear. Asking individuals about their perspectives, interests, and feelings is not something that can be easily obtained in any other way. This does not mean the data will be accurate (see Response Bias); however, a survey may still be the best way to get that data. In other situations, the need for self-report may be a matter of convenience rather than need. The decision to use a survey may apply to all the information you want or just part of the data you hope to collect. For example, you may be able to get some of the data you need in another way; doing so may provide more accurate results but, for reasons outside of your control, it may be easier to simply ask. For example, the Family Educational Rights and Privacy Act (FERPA) may prohibit you from obtaining information about individuals without specifically getting permission from each individual. Then, even if participants grant you permission, you may find it difficult to obtain the data you need from those who have it for political or practical reasons (e.g., institutions may be unwilling or unable to provide the data). It may be simpler to ask individuals to provide the data even though by doing so you may lose precision or accuracy due to some form of response bias. In either case, the primary issue that will determine the need for a survey is whether self-report data collection is required.

### Survey Need Example

Continuing with the example for counseling services, suppose you wanted to determine the relationship between a student's achievement and their perceptions of counseling services. To do this, you would need to know the current grade point average (GPA) of students, as well as their personal feelings about counseling services. You would definitely need to ask students about their feelings, but it may be best to obtain their GPA in another way. A GPA is often calculated differently depending on the school and purpose for the score; as a result, and for a variety of reasons, when asking about a student's GPA, the student may be unable to share this with you (e.g., they may not know their current GPA and provide only an estimate, or they may report the calculation based on all subjects instead of a specific set of core academic subjects you are interested in). In either case, getting GPA data from a more accurate source rather than asking students would be preferred. It may be more difficult to accomplish, but it would be more accurate. However, you may need to choose feasibility over accuracy. It may be impossible to get permission from respondents to get grade information, or the institution may not be willing to accommodate your request for the information even if you have student's permission. In such cases, if you are willing to accept the effect additional measurement error will have on the results, then asking students to self-report may be acceptable.

## Who Should Be Invited to Take the Survey? Defining the Population

Another important pre-item-development task involves deciding who you feel should be invited to take the survey. Basically, this involves defining the population and understanding basic contextual issues surrounding the population. This step in the process addresses two important factors: accuracy and access. *Accuracy* involves making sure you ask the right people. Properly defining the population helps with issues of accuracy. *Access* is a feasibility issue and addresses whether you have access to the right people. Both these factors will influence sample size adjustments and sample selection procedures.

It is essential that you ask the right people. Asking the right people means asking those who have direct knowledge and would best be able to provide accurate information. A common mistake some researchers make when administering a survey is using a convenient indirect sample (i.e., people who do not have a direct knowledge but are associated with those who do). For example, asking parents to speculate what their children might feel or do. Given ethical and feasibility considerations, it might be more convenient to ask parents to complete the survey rather than their children. However, if you ask parents to speculate on things for which they do not have a direct knowledge, their indirect conjecture will likely produce flawed results.

In addition to asking the right people, you need to consider sampling; specifically, how many participants you need to survey. The issue of sampling is addressed in a separate chapter (see Sampling); however, the main concern is getting a representative sample. You need to be confident that the responses you obtain represent the targeted population. This includes any subgroup within that population, should the study's purpose require disaggregation of data.

Getting access to those you wish to survey is also important. This is a feasibility issue and a question of whether the topic is researchable. If you cannot get reasonable access to those you plan to survey, the topic cannot be studied. This may also determine how the survey might best be administered (e.g., online, by mail, personal interview, or phone). Depending on the situation, you may decide accessing participants might best be accomplished in a specific manner and at a specific time.

## Participant Selection Example

Continuing with the example for counseling services, because the target population is first-year university students, it would not be wise to survey instructors who teach undergraduate courses asking what they think first-year undergraduate students think and feel. While this seems pretty obvious, too often similar missteps occur. In addition, you need to make sure you survey a sufficient number of first-year undergraduate students. If you plan to disaggregate results, you also need to make sure your results adequately represent, not just the entire population, but also the subgroups of interest. For example, if a specific demographic is of interest (e.g., gender, age, race, socio-economic status, or program enrollment), you need to make sure that you get a sufficient number of respondents from each subgroup. Gaining access to the participants can also be a challenge. Suppose you plan to administer an online survey; you need to have email addresses for students. Assuming they all have an email address, you also need to make sure you have the email address students typically respond to (e.g., university-issued email or personal email). In addition, you will need to decide when it is best to administer the survey. Sending out surveys during final exams may exaggerate measures of stress. Administering the survey once school has ended for the summer may minimize results and reduce response rates.

## Chapter Summary

- Identifying what data is needed will help determine how to best get that data.
- Not all studies should use a survey.
- A survey should be used only when self-report is the best way to obtain the information needed.
- Data should be obtained from a direct source.
- A researcher's ability to gain access to participants will determine whether a topic can be researched properly.

## Discussion Questions

1. When might a survey not be the best method for collecting data?
2. Under what circumstances might a study be untenable?
3. Under what circumstance would it be appropriate to survey an indirect source for information? Explain your rationale.

## Practice Tasks

1. Based on a study you might complete, identify the population you will need to target. What is the best way to access these individuals?

# Sampling Basics

## Principles and Procedures

Sampling

The results from a survey are used to describe the characteristics of individuals in a specific finite group. In cases where the population is small, it may be prudent to survey the entire group (i.e., a census not a sample). However, when large groups are targeted, a census may not be possible and likely are not needed. Carefully sampling the population allows us to make fairly accurate generalizations about the population from which the sample comes. The degree to which the generalizations being made are valid largely depends on the sampling methods used and the response rate. Basically, the validity of the conclusions will depend on whether the sample used is a representative sample, meaning those responding to the survey are a reasonable representation of the population targeted in the study.

To ensure full transparency when conducting survey research, best practice requires that you always report the sample size, the response rate, the way in which samples were selected, and if known, the population size or at least an estimate of what the population is assumed to be. In addition, you need to report issues that might have occurred in terms of any systematic response refusal patterns.

This chapter will explore various issues related to sampling and how to determine an appropriate sample size for a survey research study.

# Measurement Error Related to Sampling

With regards to sampling, there are generally two ways that measurement errors can occur. The primary problem is sampling errors, but response refusal issues can also increase measurement error. You will likely never know the degree to which these errors have affected a study. The primary way to alleviate these problems is to increase the size of the sample. Still, you should always assume there will be some degree of measurement error.

*Sampling error* occurs when those chosen to take a survey do not adequately mirror (represent) those in the population from which the sample was drawn. Fundamentally, this is an issue because the results over- or underrepresent various subgroups within the population. When this happens, the results obtained from the survey are not generalizable and are considered flawed (biased). This can easily happen with small samples, especially when there are several smaller subgroups within the population; however, survey error can also be the result of the sampling techniques used.

*Response refusal* can be another problem. Even when efforts to adequately sample the population are made, you cannot force individuals to respond to the survey. When a large number of potential respondents choose not to complete a survey, the results may not be generalizable (i.e., they may not accurately represent those in the population). The degree to which response refusal affects generalizability depends on whether the refusal pattern is random or systematic. When potential respondents do not complete the survey, you need to determine whether the response refusal is spread equally across the sample regardless of an individual's characteristics or subgroup membership. This

would mean the response refusal is random, in which case the sample will continue to be fairly representative of the population. However, if the response refusal is systematic—meaning one group of individuals with similar characteristics is more likely not to complete the survey compared to other groups of individuals—then you have a problem.

## Oversampling as a Solution

One common solution to both these problems is to increase the sample size by oversampling the population. With random response refusal and sampling error in general, oversampling may help solve the problem. However, increasing the sample size will not fix the problem of systematic response refusal. For this reason, you should always examine the characteristics of respondents who choose not to take the survey to determine if any discernable pattern can be found and to make sure there are no systematic response refusal issues. This may not be possible if you do not have access to information about non-respondents.

It is generally a good idea, at least in theory, to use as large a sample as possible. This will maximize the likelihood of obtaining good representation for the general population and subgroups within that population. In practice, if you are studying a small, finite population, you may need to invite all to participate. However, if the population is large, you may not be able to survey everyone, and you likely don't need to. One argument for not increasing the sample size is cost, but with the prevalence of online surveys, cost is often not an issue. Access can be a challenge and a good reason for not surveying the entire population (e.g., when the population includes younger children). Perhaps the best reason for not obtaining a larger sample is that it simply is not needed. At some point, surveying additional individuals would not change the result, and oversampling may lead to survey fatigue (and decreased response rates). If those in a particular population are inundated with invitations to complete surveys, they are less likely to respond. This can negatively affect everyone trying to obtain survey data from that population. Survey fatigue associated with receiving frequent invitations to complete surveys is a serious concern for researchers.

## Incentives and Compensation

One way some choose to alleviate the survey error and response refusal problems—one that may also reduce systematic response refusal issues—is to offer respondents an incentive to complete the survey. This can be done by offering payment or some other form of compensation. Some surveys offer the opportunity to win a reward (e.g., offering those who complete the survey entry into a prize drawing). This has been found to be effective in some cases but not all. The degree to which compensation will be effective depends on how enticing the incentive is for the participants and the integrity of the individuals. In truth, offering incentives in not always effective and may have some unintended negative consequences. Certainly, adequately compensating people for the time and effort it takes to complete a survey may increase response rates and reduce systematic error. However, while more respondents may complete the survey, several might do so without any intention of answering items thoughtfully or accurately; they simply want the compensation. Likewise, if the incentive is somehow coercive (individuals are forced to complete the survey), not only does this potentially violate ethical practice guidelines but it will also likely produce flawed data if respondents don't take the time to reflect and respond honestly. If the research requires Institutional Review Board (IRB) approval, further guidelines for offering incentives may also apply.

## Determining an Appropriate Sample Size

To get a rough estimate of the number of individuals that should be included in a sample (i.e., a sample size large enough to accurately represent the population), you can use previously calculated estimates, organized and provided in statistical tables like the one below. There are also online sample size calculators. Looking at Table 1, you will note that when the population you are trying to describe is small, you will need to survey a large proportion of the population in

order to have confidence in your results. On the other hand, once your population gets to a certain size, increasing the size of the sample will probably not improve the results.

Based on these data, a sample size of 300 to 400 will likely suffice in most cases. For populations less than 1000 individuals, you will need to obtain responses from a large proportion of the population and, if feasible, you may wish to invite all those in the population to respond for response refusal reasons. For small populations of 100 or less, you will likely need to survey almost all those in the population if you wish to obtain valid results.

**Table 1**

*Recommended Sample Size for a Given Population*

| Estimated Population | Required Sample Size | Estimated Population | Required Sample Size |
|---|---|---|---|
| 500,000,000 | 384 | 1,000 | 278 |
| 10,000,000 | 384 | 750 | 254 |
| 500,000 | 384 | 500 | 217 |
| 100,000 | 384 | 400 | 196 |
| 75,000 | 382 | 200 | 132 |
| 10,000 | 370 | 100 | 80 |
| 5,000 | 357 | 50 | 44 |
| 3,000 | 341 | 25 | 24 |
| 1,500 | 306 | 10 | 10 |
| Adapted from Krejcie, and Morgan, (1970) | | | |

*Note*. Recommended sample size is based on a required 95% confidence level.

## Factors that Affect Sample Size Requirements

The basic sample size estimates may need to be adjusted given the purposes for the study and data analysis needs.

- The more homogeneous (i.e., similar or like-minded) the population, the smaller the sample size can be.
- The better the sampling procedures, the smaller the sample can be (noting that some sampling techniques require larger samples to reduce measurement error).
- The more planned comparison breakdowns (i.e., disaggregation of data based on multiple characteristics), the larger the sample needs to be. For example, you may wish to report comparisons between two groups of respondents, with an additional breakdown for each group based on gender and/or various age categories. A larger sample is needed to ensure sufficient numbers in each breakdown category.
- If the survey is to provide data that will be analyzed using sophisticated statistical procedures, additional respondents may be needed to satisfy the requirements of that specific procedure (e.g., a multiple regression or confirmatory factor analysis).
- The sample also needs to be larger if there is a greater likelihood of response refusal. Once you have determined a sample size that will likely produce a representative sample, the formula to calculate a revised estimate that accounts for response refusal requires you to estimate the proportion of invited participants likely to actually respond.

$$Adjusted\ Sample\ Size = \frac{Estimated\ Sample\ Size}{Proportion\ Likely\ to\ Respond}$$

# Calculating Required Sample Sizes

Using tables to estimate an appropriate sample size is based on the assumption that the sample needs only be sufficient to provide a representative sample for the population. These approximations also assume you know or can estimate the population size. Using tables to determine sample sizes can provide a reasonable estimate, especially when the survey is designed to capture several variables. However, you may need to calculate a sample size for a specific variable based on a particular level of precision, confidence, and variability. In these cases, the sample size calculation will depend on a few considerations (criteria).

## Level of Precision (e)

Sometimes called sampling error, the level of precision indicates how accurate you want the result to be. For example, the results may need to be within a specific confidence interval (e.g., within ± 5% or within a 95% confidence interval). In this case, we would expect the true value (population parameter) to be within the specified range around the statistic obtained from the survey. In this case, if the precision needed to be ± 5% (a common value because it represents a 95% confidence interval), we would set e = 0.05 to indicate that level of precision.

## Risk Level or Confidence Level (z)

The confidence level is an indication of the risk you are willing to accept that the statistic (i.e., the mean or proportion being measured in the survey) is within a specific distance from the actual population parameter. The risk level (z) is based on the Central Limit Theorem which proves that the mean values of repeated samples drawn from a population are normally distributed. In a normal distribution, we know that 95% of the sample mean values, obtained from repeated samplings, will fall within 1.96 standard deviations of the population's actual mean value (i.e., the population parameter). With this level of confidence, there is only a 5% probability that the sample mean values you obtained will be extreme (out of the ordinary, far from the actual population parameter). In order to lower the risk of getting an extremely erroneous estimate of the population parameter, you would need to choose a higher confidence level (z). For example, if you wished to lower your risk you could set the confidence level to 99% and set the z = 2.58 to indicate that level of risk. Using a larger confidence level will result in a larger sample size estimate, which lowers the risk that the sample will produce an extreme outlier.

**Table 2**

*Common Risk Level Values*

| Confidence Level | Critical Value (z) |
|---|---|
| 80% | 1.29 |
| 90% | 1.65 |
| 95% | 1.96 |
| 99% | 2.58 |

## Degree of Variability (*p*)

The variability level is an indication of the expected difference in response values, or prevalence of individuals with a specific characteristic within the population. Variability will range from 0 to 1 with a $p = .50$ representing maximum variability. The more homogenous (similar) the population the less variability there will be in responses. The more heterogeneous (dissimilar) the population, the more variability exists, which requires a larger sample size to obtain a generalizable result. Often, we don't know the amount of variability that is likely to exist. Other times we might anticipate the variability from previous studies or antidotal observations. If the variability is unknown, a common practice is to set $p = .50$ (i.e., maximum variability). This is not necessarily best practice; however, using $p = .50$ will produce a conservative (larger) sample size estimate due to the expected dissimilarity of individuals in the population.

Determining the degree of variability is complicated when proportions are not a dichotomous condition. For example, the degree of variability for individuals with red hair can be estimated from previous approximations of those in the population. If 20% of those in the population tend to have red hair, then $p = .20$. However, when the condition being studied is not a dichotomous choice, determining variability is more difficult. An example of this is the case where you ask people whether they agree with some statement. Often surveys use Likert scales (e.g., *strongly agree*, *agree*, *disagree*, *strongly disagree*) to capture information. Variability in this case is not a dichotomous condition. You can set the variability based on a single condition (e.g., those who strongly agree) but that leaves out those who agree but not strongly. You might need to collapse categories to combine similar conditions (e.g., those who agree or strongly agree). In either case, $p$ should represent the expected prevalence of the condition in question.

When an item on a survey is used to measure something that is not a proportion but rather a scale or continuous value, the variability estimates require we use estimates for the variability of the mean. For example, a survey may be used to ascertain the ages or height of individuals in a population, which are both continuous variables. In these cases, $p$ represents the expected variability of the mean, not the number (i.e., proportion) of individuals exhibiting a specific condition or characteristic. This is a challenge because we often do not have a good estimate of the population variance. Furthermore, there are often multiple values being obtained from a single survey; depending on which value is used, the sample size determinations can vary widely. If variance is unknown (and cannot be easily estimated), or if there are several continuous- and proportion-based variables being measured in the survey, then the sample size calculation obtained is often simply a guess. For this reason, basing sample size requirement on these types of variables is a challenge and often avoided. Cochran (1977) however suggests you might estimate the variability of a continuous value using one of the following methods.

- Pre-sample the population to obtain an estimate.
- Use values obtained when pilot testing the instrument.
- Use variance obtained from previous studies.
- Make an educated guess based on what you know about the population.

## Definitions

Before we consider various ways to calculate potential sample sizes, we should review a few definitions.

A **statistic** is a value obtained from a sample.

A **parameter** is a value obtained from a population.

For example, a survey may be used to determine the proportion of individuals who indicate agreement with a specific stance. Individuals either agree or disagree. If the result is obtained from a sample, it is called a statistic. If the result is obtained from the population, it is called a parameter (sometime referred to as the true value, assuming no measurement error).

### Definitions for Variables Required to Complete Sample Size Calculations

**Sample size (n)** is the estimated size of the sample required to obtain an adequate estimation of the population parameters.

**Population size (N)** is the size (or estimated size) of the population.

**Confidence Level (z)** indicates the risk you are willing to accept that the statistic obtained from a sample will not be very close to the actual population parameter. Using z = 1.96 would indicate a 95% level of confidence is required.

**Degree of Variability (*p*)** indicates the response variance you expect to obtain from individuals in the population. Using *p* = .50 would indicate the maximum amount of variability is likely to occur.

**Level of Precision (e)** indicates the amount of sampling error that would be acceptable. Using e = 0.05 indicates you expect the true value (parameter) for the population to be within ± 5% of the statistic obtained from the selected sample.

## Basic Sample Size Formula (Known Population Size)

The allure of this approach is that you need only two pieces of information—the population size and the desired level of precision (see Yamane, 1967).

$$n = \frac{N}{1 + N(e^2)}$$

## Calculating Sample Size (Proportions)

This calculation is used in situations where a single item from a survey is intended to provide context. The formula uses an estimate for the proportion of individuals who exhibit a specific characteristic or attribute as the basis for the degree of variability expected (see Cochran, 1977; Daniels, 2018; Israel, 1992). For this calculation, the expected degree of variability is considered in addition to the confidence level and level of precision required but does not require the population size to be known.

$$n = \frac{z^2 p(1-p)}{e^2}$$

An adjustment for smaller finite populations is possible if the size of the population is known.

$$n_o = \frac{n}{1 + \frac{(n-1)}{N}}$$

## Calculating Sample Size (Scale/Continuous Variables)

One formula that can be used for scale or continuous values employs variance of the mean ($\sigma^2$) instead of variability based on proportions, *p*(1-*p*). Because a good estimate of the population variance is often unavailable, determining sample size using estimates of variability that are based on proportions is frequently preferred (Cochran, 1977).

$$n = \frac{z^2\sigma^2}{e^2}$$

## Determining Sample Size Example

Returning to the counseling service example from previous chapters, now that we have conceptualized the study, suppose you now want to decide how many first-year undergraduate university students should be surveyed. It is likely that you know how many students are enrolled (say N=5000). Using a table of suggested sample sizes, you would get your answer, n = 357. This number is based on a 95% confidence interval. Using a simplified formula based on the population mean and level of risk (Yamane, 1967), we would get a similar answer, n = 371.

$$n = \frac{N}{1 + N(e^2)} = \frac{5000}{1 + 5000(.05^2)} = 371$$

Alternatively, you may wish to get a second opinion because you want to determine the proportion of students who exhibit symptoms of depression. With the expectation of a 95% confidence interval, 5% risk level, and based on previous estimates that suggest 40% of students typically suffer from depression, you would get an n = 369.

$$n = \frac{z^2p(1-p)}{e^2} = \frac{1.96^2(.4)(.6)}{.5^2} = 369$$

Since you know how many students are enrolled, we can adjust this estimate to reflect the population size. Noting however that this may not be appropriate as the population size isn't really that small.

$$n_o = \frac{n}{1 + \frac{(n-1)}{N}} = \frac{369}{1 + \frac{(369-1)}{5000}} = 344$$

These values are all fairly similar, but it is likely that not all those invited to participate will complete the survey. In fact, from previous experience suppose you believe that only about 25% of those invited to participate will actually complete the survey. Accounting for this, it is reasonable to assume that you will need to send the survey out to over 1,376 students if you hope to obtain the number of responses you need.

$$Adjusted\ Sample\ Size = \frac{Estimated\ Sample\ Size}{Proportion\ Likely\ to\ Respond} = \frac{344}{.25} = 1376$$

## Reflection Exercise

Reflect and be prepared to discuss the following questions after reviewing the sample size example presented above, noting the topic being addressed (prevalence of depression among first-year undergraduates).

How likely do you feel a systematic response refusal pattern might develop? Explain.

Suppose it is likely that the refusal to complete the survey pattern is systematic. Would increasing the sample size help? Why or why not?

What are the benefits and potential limitations for incentivizing participation?

## Chapter Summary

- Not all survey research requires sampling. With smaller populations, a census (surveying the entire population) is required. With large populations, a properly selected sample will negate the need for a census.
- When sampling is the best course of action, those in the sample need to adequately represent those in the population. This is called a representative sample.
- Regardless of the sampling techniques used, some sampling error will inevitably occur.
- Sampling error occurs when the sample does not adequately represent the population.
- In addition to inadequate sampling procedures, response refusal can affect whether a sample is a representative sample.
- There are several ways to estimate the sample sized needed to obtain a representative sample; however, several additional factors will influence the required sample size—including characteristics of population and data analysis needs.
- Increasing the size of the sample is almost always preferred to alleviate issues of sampling error and response refusal. However, there are times when getting a larger sample may not be feasible or cost effective.

## Discussion Questions

1. Explain the benefits and disadvantages of using a sample.
2. Explain how survey fatigue affects response refusal. How do these issues affect sampling?
3. Explain why oversampling likely will not solve systematic response refusal issues.
4. Does setting a lower risk level guarantee the statistic you obtain will equal the population parameter? Explain.

# References

Cochran WG (1977). Sampling Techniques, 3rd edition. New York: John Wiley & Sons Link to ebook

Daniel, W. W., & Cross, C. L. (2018). *Biostatistics: a foundation for analysis in the health sciences*. Wiley.

Israel, G. D. (1992). Determining sample size. Link to Article

Krejcie, R.V. and Morgan, D.W. (1970) Determining Sample Size for Research Activities. *Educational and Psychological Measurement, 30*, 607-610.

Yamane, T. (1967). Statistics, An Introductory Analysis, 2nd Ed., New York: Harper and Row.

# Sampling Techniques and Procedures

Sampling

As previously mentioned, there are many reasons why you would use a sample rather than a census when conducting research. And as mentioned, there are many things that could go wrong. One of the things that could go wrong is the selection of a sample. The primary goal of sampling is to create a representative sample, one in which the smaller group (sample) accurately represents the characteristics of the larger group (population). If the sample is well selected, the sample will be generalizable to the population.

There are many ways to obtain a sample. The techniques used will vary based on the circumstances under which the study is conducted as well as the aims of the research. The way in which samples are drawn will affect the quality of a study.

Prior to choosing a selection method, you should have defined the population and the purpose for the study. Clearly defining the target population is important, meaning you will define both the size of the population and the accessibility of the population. As noted previously, anticipated survey response refusal will affect the size of the sample needed. Accessibility of individuals within the population will also affect the sample selection procedures. There are two general approaches to sampling: random and non-random. However, additional consideration should be made based on whether the study will be a qualitative study.

## Definitions

**Quantitative sampling.** Surveys are typically designed to produce descriptive numerical statistics (e.g., scores, ages, strength of opinions, frequencies) that can be used to describe various characteristics found within the population. Any qualitative data obtained is typical categorized and quantified. Sampling for these studies must produce representative samples because generalizability is important. A distinctive aim of these studies is to gain a general understanding of the characteristics found in the population.

**Qualitative sampling**. Qualitative studies are not interested as much with generalizability as they are with understanding a phenomenon. The sample must produce good informants for the study. The characteristics of the respondents are more important than the size of the sample. These samples will be smaller and less representative but should provide researchers access to a good representation of key informants. Often the aim of qualitative research is to get a deeper, fuller understanding of the topic or phenomenon.

**Random sampling**. A selection technique where every unit in the population has an equal chance of being selected. The unit of analysis often involves individuals but may be intact groups.

**Non-random sampling**. While random sampling may be preferred, there are many ways in which a planned random sample may become less random. In non-random sampling (or non-probability sampling), researchers are unable to select participants at random from the population. This includes situations where circumstances (e.g., survey refusal leading to low response rates or missing contact information) diminish the likelihood that the sample provides a good representation of the population. Follow-up contact or a post-survey examination of demographic characteristics are often needed to verify the degree to which survey results might be considered generalizable.

**Random selection and random assignment.** These two terms should not be confused. Random selection is used to establish a sample. If done properly, the results of the study are believed to be generalizable. Random assignment is use in experimental studies. Randomly assigning individuals to two different groups is done in an attempt to make the two groups comparable. Random selection affects claims of generalizability. Random assignment is the basis for experimental claims of causality.

# Random Sampling Techniques

## Simple Random

For this type of sampling, each individual (or unit) in the population has an equal and independent chance of being selected. In probability sampling, another name for random sampling, the researcher can select the level of chance. In order to produce a true random sample, the population must be known. A known, finite population is one where all members of the population can be identified and are accessible. This kind of sampling also assumes that all who are selected to be part of the sample will respond.

A random sample does not guarantee that the sample will properly represent those in the population. Sampling is not a precise science. There is still a chance that a randomly selected sample will be skewed in some way—by this I mean the sample under- or overrepresents some group or characteristic found within the population. The Central Limit Theory tells us that when an infinite number of samples are taken, the distribution of the sample means will be normally distributed, and the average of the sample means will be that of the population. However, our understanding of the normal curve likewise indicates that the mean of any one sample may be extremely different from the population. Still, while the result we obtain will not be perfect, care should be taken to attain the best result possible. Random sampling

is used when we don't have specific information about those in the target population and wish to remove human bias from the selection process. Random sampling is believed to be the best way to avoid selection bias.

## Systematic Sampling

Systematic sampling is an adaptation of random sampling which does not give everyone an independent chance of being selected. For example, the selection process may choose every fifth person in a list. This is not completely random because an individual's position in the list limits the chance they will be selected (the selection is dependent on the individual's position in the list); the randomness of the selection becomes even more problematic if the list is compiled in a way that introduces a systematic bias.

## Stratified Sampling

A *stratum* is an identifiable, mutually exclusive subgroup within a population. Stratified sampling attempts to guarantee representation from each important strata within the population. Membership in a stratum must be homogeneous so the sampling would not allow selection of an individual who has membership in two distinct strata. Stratification is considered to be a random sampling technique because individuals are randomly selected from each stratum. Stratified sampling could be equal or proportional. The researcher could select an equal number of participants from each stratum, or they could select participants proportionally based on the estimated size of each stratum. Proportional sampling is preferred if the sample is to be generalizable. In this case the required sample size selected from each stratum should be determined independently so each stratum is appropriately represented. This may require a much larger number of participants compared to the number needed using simple random sampling.

## Cluster Sampling

With cluster sampling the unit of analysis is based on intact groups rather than individuals. For example, all those in a particular school or classroom are selected, not specific individuals within each school or classroom. The intact units are however randomly selected. For this to produce a representative sample, it is assumed that the intact units will include a variety of individuals represented in the population or that an adequate number of heterogeneous intact groups selected will, as a whole, adequately represent the population. This may or may not be the case and may require a combination of stratified and cluster sampling. In practice, not all samples obtained in this manner are random samples. When a research study requires that the unit of analysis includes sampling of intact groups, special care needs to be taken to make sure that adequate representation is obtained.

# Non-Random Sampling Techniques

While random sampling is preferred (and considered by some to be the gold standard), it is not always possible to obtain a random sample. And while the basic procedures used with non-random sampling often mirror sampling procedures used to obtain random samples, any method of sampling that does not allow for individuals (or units) to have an equal and independent chance of being selected is referred to as non-random sampling. Non-random sampling is considered inferior to random sampling because there is a greater chance that the sample will not represent the population adequately. However, for a variety of reasons, non-random sampling in the social sciences is quite common.

The most common reason for using non-random sampling is that of necessity. Random samples cannot be selected when the size of the population is unknown, individuals cannot be easily identified, access to the potential respondents is restricted, or contact information is unattainable. In addition, even when random selection is implemented, ethical consideration regarding the protection of human subjects' rights may prevent the sample from being a true random sample. For example, randomly selected individuals may not be willing to provide information or allow their information to be used. If this happens in large numbers, or in a systematically unbalanced way, a potential random sample will, in practice, become a non-random sample. This could considerably diminish the chances that the sample adequately represents the population.

Regardless of the way a sample is obtained, the goal of any sampling technique is to allow the researcher to access information from those who can provide useful information. Useful in this case means providing information that helps answer the research questions in such a way that researchers can trust the results; this is an issue of validity. There are many ways to obtain a non-random sample.

## Convenient Sampling

A convenient sample is comprised of individuals who are available and willing to complete the survey (i.e., volunteers who can be contacted and are willing to participate). Any time you send out a broad invitation to potential respondents asking them to volunteer to take a survey, you are creating a convenient sample. A convenient sample is less likely to adequately represent the population than a random sample, and the results are less likely to be generalizable without having a larger sample size. Even when a high response rate is obtained, if those available and willing to participate systematically do not represent those in the population, the results will not be valid. Unfortunately, we may never know the degree to which any sample is biased, but there is an increased probability that a convenient sample will not adequately represent the population compared to a random sample.

## Quota Sampling

Like a stratified sample, quota sampling involves selecting individuals to participate based on identifiable characteristics of individuals within the population. With quota sampling, the researcher identifies major subgroups of interest within the population (strata), determines the number of individuals needed, and then attempts to obtain a sufficient number of willing and available participants from each subgroup. Like stratified sampling, the number of participants needed (i.e., the quota) may be based on equal or proportional requirements.

## Purposive Sampling

Selecting participants using purposive sampling procedures requires the researcher to specify criterion for inclusion. As a result, purposive sampling has at times been called criterion-based sampling. Criteria are based on a set of characteristics individuals possess (i.e., things about the potential respondents that make them interesting because they would likely be able to provide useful information). Once the criterion for inclusion have been identified, participant selection will focus on getting a sufficient number of willing participants who meet the criterion. Because participation in a qualitative study often requires participants to willing submit to lengthy, involved data collection procedures, the sampling techniques used in qualitative studies are almost always purposive. There are several ways the inclusion criterion for purposive sampling may be established.

## Comprehensive Sampling

Comprehensive sampling attempts to obtain data from individuals experiencing every possible condition or subgroups defined within the population. This usually isn't possible, but when it is possible, it is not practical. More often researchers will use some form of homogeneous sampling where selection criteria are based on choosing individuals with similar experiences, situations, perspectives, interests, or circumstances. This is a more manageable approach and researchers often will refine inclusion criteria to match a particular research purpose. Following are examples of these inclusion criteria.

**Maximum Variation (Intensity Sampling).** Selection criteria are designed to obtain a wide range of participants based on a few specific variables. An example of maximum variation in a sample would be the selection of students with various levels of academic achievement from various years in school.

**Extreme Case.** In this situation, selection criteria are intended to include participants representing extreme situations. For example, those who participate in a regimented exercise routine every day without fail and those who claim to never exercise at all.

**Typical Case.** With this strategy, the researcher sets inclusion criteria to include people who typify the normal (most prominent) individuals in the population. To do this the researcher would consult experts or examine theory to

determine characteristics of the “typical” person they wish to study, then set out to find a sample of these individuals. For example, the researcher might look for individuals described as being typical based on characteristics like age, experience, education, gender, behavior, or perspective. In cases where the purpose of the research is to define what is typical or normal, the sampling would need to be more comprehensive.

**Critical Case.** Sampling to include critical case individuals requires identifying individuals or intact groups who are important for some specific reason. For example, a researcher might select schools where conditions would likely result in greater resistance to planned reforms. The critical inclusion requirement being that if there is resistance, it will exist in those schools. The converse may also be a critical case; if there is little resistance to the proposed reforms, it will likely be at other schools.

**Negative Case (Discrepancy Sampling).** The selection criteria for a negative case are intended to identify respondents who are atypical, go against the norm, or provide examples that might disconfirm expected results. The sample is chosen to include those who appear to wholly disprove or refute a theory. For example, an intervention may be extremely effective for the vast majority of individuals; however, a small group of individuals tend to be negatively impacted by the intervention, meaning those individuals represent a negative case by going against expected outcomes.

**Referral or Snowball Sampling.** This type of sampling is based on practical purposes rather than research purposes. When those individuals matching a particular set of criteria are not readily identifiable, one way to locate participants is to ask for referrals. Once one individual is found and surveyed, they are asked if they know others who share similar characteristics. Thus, the selection process has a snowball effect (i.e., the sample gets larger as you go). With this technique it can be difficult to know when the number in the sample is sufficient. This is where data saturation decisions need to be made. *Data saturation* refers to situations where the information you obtain from participants begins to repeat. Saturation refers to the point where you don’t need more participant data because you are getting the same answers. Additional information in this case would not improve your understanding of the phenomenon, just substantiate the strength of the finding.

# Need for Replication

It is important to understand the unlikelihood that any sample you obtain will perfectly represent the population from which it was drawn. Even with a random sample, there is a high probability that the sample will not exactly represent the population in some way. Given that sampling is not a precise science, the need to replicate a study should be evident. A carefully selected sample can provide valuable results, which is why we conduct research. However, the sample used to obtain a result may have been flawed in some way, thus you would need to redo the study with a different sample. In this sense, replication of a study is done to verify the results. Still, few studies are replicated in such a way that completely verify the results of previous studies. The outcomes obtained from any carefully constructed sample will likely be of some value, they just won’t be perfect.

## Chapter Summary

- The way in which a sample is obtained will affect the quality (or value) of the sample.
- Quantitative surveys are typically designed to produce descriptive numerical statistics that can be used to describe various general characteristics found within the population.
- Qualitative surveys are not interested as much with generalizability as they are with understanding a phenomenon. As a result, in a qualitative study, the sample must produce a good set (sample) of informants for the study.
- Random sampling refers to sampling techniques that allow an equal and independent chance for participants to be selected.
- Random sampling is believed to be the best way to alleviate sampling error. Sampling error affects the degree to which the sample represents those in the population and thus the generalizability of the results.
- Random sampling is not a foolproof method, and any random sample has a high probability of being flawed in some way. Major flaws in the sample obtained have the potential to adversely affect the result. Minor flaws can be acknowledged and accounted for.
- The ability to produce a true random sample will be dependent on whether the size of the population is known (finite), individuals can be easily identified, access to the potential respondents is unrestricted, and the contact information for potential participants is available. In addition, ethical consideration regarding the protection of human subjects' rights and response refusal issues may prevent a true random sample from being obtained.
- The most convincing reason for using random sampling is that it helps researchers avoid human bias in the selection process. Random sampling is often used when specific demographic and personal information about individual respondents is unavailable.
- Non-random sampling is commonly used in the social sciences due to the difficulties in obtaining a true random sample.
- Random sampling and non-random sampling techniques are similar with the exception of random selection.
- The most common reason for using non-random sampling is one of necessity—there is no other way to proceed.
- Convenient samples are based on participants' willingness and availability (i.e., volunteers).
- Other selection parameters can help refine the sampling procedure (e.g., purposive sampling). This is only possible when specific attributes of the potential respondents are known.
- The research purpose will often dictate the best sampling techniques to use; however, practical issues will also influence the decision.
- The unit of analysis is often individual people in the population; however, sometimes intact groups are selected. When intact groups are used (i.e., cluster sampling) the degree to which adequate representation has been achieved must be carefully considered.
- Because no sample is perfect, replication studies are useful to validate the results of any study.

## Discussion Questions

1. What are the benefits of using a random sampling procedure over a non-random sampling procedure?
2. How likely is it that any sample you select will be perfect? Explain.
3. What is the purpose of a replication study? When and why are they needed?

## Practice Tasks

1. Pretend you wish to make comparisons between specific groups of individuals within a population. What sampling techniques would best serve your needs? Explain the benefits and limitations of the sampling procedures you chose.
2. For a specific study you might consider completing, identify the population and chose a sampling technique that would serve your needs. Explain the benefits and limitations of the sampling procedures you chose. What particular challenges will you need to overcome in order to obtain the sample?

# References

Davies, R. S., Williams, D. D., & Yanchar, S. (2008). The use of randomisation in educational research and evaluation: A critical analysis of underlying assumptions. *Evaluation & Research in Education*, *21*(4), 303-317.

# Survey Error and Response Bias Problems

Survey Research

## Mistakes Were Made!

The most common mistake many researchers make when conducting survey research is administering the survey before it has been properly vetted. And unfortunately, another common practice that fuels the criticism of survey use is hiding a survey's flaws—treating the results as adequate when they are not. While it is unlikely that any survey you might design will be perfect, understanding and attempting to mitigate potential problems will increase the usefulness of any survey data you obtain.

## Survey Errors and Response Bias

There are many ways in which a survey research study can go wrong. When a survey goes wrong, we describe the results as having error or bias. Both can invalidate the results of a study. *Survey error* is caused when researchers make mistakes when creating or administering a survey, as well as when they interpret the results of a survey. *Response bias* refers to the ways respondents may be unduly influenced when providing answers on a survey. Bias is an issue that affects the accuracy of the survey data obtained and is the result of participants' inability or unwillingness to answer questions precisely or truthfully. There are several specific kinds of error and bias that are important to understand prior to developing your survey.

### Survey Error

**Survey error** refers to mistakes made in the construction and implementation of the survey instrument, as well as the interpretation of results. Survey error is associated with completeness, interpretability of the data, and the generalizability of the results. This type of error affects the validity of the measurement and what is called measurement error. Survey error is avoided, to some extent, by carefully designing the study, creating a comprehensive set of well-crafted survey items, and properly interpreting and presenting the results.

### Common Forms of Survey Error

**Survey Scope Error.** This error refers to the mistake some make when a survey does not include important items required to fully answer the research questions. Failing to ask important questions will provide an incomplete or inaccurate answer to the research questions posed in the study. This can be difficult, and there will always be some tension between making sure to include all essential survey questions and limiting the length of the survey. Sometimes this error is made because the researcher has not carefully considered what they needed to know. Other times you may

only realize a mistake was made once you have the survey results back and find you failed to ask an important question. When a questionnaire or scale does not include items that cover all the important aspects of the topic or construct, survey scope error occurs, reducing the validity of the results.

**Purpose Creep Error.** At times researchers (or clients) will add items to a survey that don't directly align with the needs and purposes of the study. This is referred to as purpose creep. Unnecessary survey items, like superfluous demographics or items that "might be nice to know," may be interesting, but often they are never used. Adding unnecessary or tangentially related items may not affect the validity of the results, but it can still be a problem because it can add to participants' fatigue and can affect their willingness to thoughtfully answer all the items in a lengthy survey.

**Sampling Error.** This occurs when mistakes are made selecting a sample. Sampling error typically occurs when a specific subgroup within the population is under- or overrepresented in the sample. When this happens, the results obtained from the survey are not generalizable. This is because the characteristics of respondents in the sample do not match proportionally with the characteristics of those in the population. This can easily happen with small samples, especially when there are a number of small subgroups within the population. As a result, it is generally a good idea to over-sample when possible to maximize the likelihood of obtaining a good representation of the general population and subgroups within that population. However, having a large sample size may not solve the problem if the sample does not proportionally mirror that of the population; in this case, sampling error will still cause the results to be flawed. You will likely never know the degree to which sampling error has affected a study. You should always assume there will be some sampling error and take steps to alleviate the problem as much as possible.

**Response Rate Issues (Response Refusal).** This is related to sampling error, in that it can affect the generalizability of the results, but it is slightly different. If the group of respondents you invite to take your survey would likely form a representative sample of the population, but several individuals choose not to respond to the survey, the results may not be generalizable. To ensure full ethical transparency, you should always report the sample size and the response rate when writing up results. You should also examine the pattern of respondents who chose not to take the survey to determine if any discernable pattern can be found. If the pattern is random, then there may not be a problem. If the pattern is systematic, meaning one group of potential respondents with similar characteristics is more likely not to complete the survey than another group, then you have a problem.

**Item Nonresponse Error.** This happens when a participant fails to answer all the questions on a survey. It can happen by accident or intentionally. Nonresponse is unfortunate when you are administering a questionnaire but does not necessarily invalidate the results; the results depend on how many respondents failed to answer a specific question. However, item nonresponse is definitely a problem when administering a scale. This is because the items in the scale work together as a group to provide a measure of a construct. In a scale, each item represents a specific aspect of the construct, and the measure, or score, relies on all the items being answered truthfully and accurately.

## Response Bias

**Response bias** is a general term that describes the ways a respondent may be influenced when self-reporting their answers on a survey. Bias affects the accuracy and reliability of the results. There are several ways in which an individual's response may be inaccurate. The inaccuracy, or bias, may be deliberate or subconscious on the part of the respondent; it is the result of a respondent being unduly influenced to respond in a certain way, or a respondent's unwillingness to answer the questions honestly. Bias can be caused when items are unclear or poorly constructed but can also occur as a result of the response scale selected for participants to use when they provide their answers. Never underestimate the propensity of individuals to misunderstand what is being asked or otherwise provide information that is inaccurate to some degree.

## Common Forms of Response Bias

**Recall Bias.** This is common in self-report situations when respondents are asked to provide information retrospectively. Human memory is imperfect. Some information (i.e., perspective and feeling) is more likely to be remembered than others. A person's ability to recall events and feelings will depend on the metacognitive ability of the individual and the significance of the event to that particular person. Recall often depends on the time interval between the event and when the individual is asked to recall their perceptions. A person may have forgotten the event altogether; they may remember incorrectly, or they might revise their recollection (see prestige bias).

**Social Desirability & Conformity Bias.** It can be hard for respondents to openly express non-conformity when asked to self-report their behavior, beliefs, and opinions; this is especially true when the respondent believes they may be ridiculed or despised. In such cases, respondents tend to provide a socially acceptable response (sometime subconsciously) over their true feelings. For example, a respondent may tend to agree with a statement more strongly than how they truly feel when the item addresses something that is generally seen in society as desirable or expected.

**Prestige Bias.** This bias is related to social desirability bias as it is based on an individual's personal desire to be seen in a positive light. This bias is based on personal feelings, not a general instinct for conformity. For example, respondents may round up their income or report exaggerated amounts of time spent on worthy endeavors (noting the reverse would be true for endeavors the individual feels may diminish how they are perceived). This may not involve outright lying, rather the individual may actually remember the facts inaccurately. Respondents often tend to view or recall their own situation in a more favorable light than is actually the case—subconsciously protecting their self-image or inflating their ego. It is often good practice to assume that, if a question has a potential prestige component, the responses are likely inflated to present the respondents in a more favorable light. Exactly how much they are inflated will depend on the question, context, and respondents.

**Acquiescence or Agreement Bias.** This bias is like conformity bias. However, unlike conformity bias, in this case the respondent will, in general and inadvertently, agree with statements. With this bias, participants tend to select a positive response option or disproportionately indicate a positive connotation. This bias will skew results towards the positive.

**Item and Option Order Effect Bias.** Order bias can be the result of both item order and response option order. The order in which survey items are presented can affect a respondent's answers due to a priming effect. People tend to contextualize their responses. Because of this, survey questions that come just before a particular query may provide information that respondents will use as context in formulating their subsequent answers. If a different primer was presented, the responses may be significantly different.

Two common response biases associated with response option order are primacy and recency bias. *Primacy bias* is the tendency for respondents to pick one of the first options presented to them. This can happen when a respondent quickly reads through the survey and picks one of the first response options they agree with. *Recency bias* is the tendency to pick an answer option presented at the end of a list. For example, in a long list of options, the choices respondents read last are more memorable to the respondent as they select an answer.

**Mood Bias and Emotional Mind Sets.** One's mood or mindset will affect the way responses are provided. For example, if a participant is exceptionally happy or angry for some reason while taking a survey, their emotional state affects the general pattern of responses provided. Given time, the respondent's current extreme emotions may subside, which will modify the intensity of the responses provided. Emotional responses can be intense in either a positive or negative direction. You will also see this when the survey addresses an emotionally charged topic. Responses may tend to be on the extreme ends of the response scale, possibly because those who choose to complete the survey have strong opinions; however, mood bias becomes a problem when opinions are exaggerated by the respondent's current emotional state.

**Central Tendency Bias.** This bias refers to the tendency of some individuals to avoid responding in extreme ways. For example, some people may never indicate they strongly agree or are extremely dissatisfied (i.e., nothing is perfect, and

nothing is completely without merit). This is the opposite of a mood bias in that responses from those who have this bias will trend closer to the center of the response scale.

**Demand Characteristic Bias.** A demand characteristic is used to describe specific cues in research that may inadvertently influence a participant's response. A demand characteristic can manifest in a number of different ways if the researcher is not careful when designing and proceeding with a study. In social science research, demand characteristics can create bias when the subject becomes aware of the purpose of the study. This may potentially bias or invalidate the outcomes. When a respondent becomes aware of the reason or purpose of the study, they may intentionally provide answers they feel would influence the results. For example, if a respondent figured out that the results of a survey will be used to set policy, the individual may attempt to answer in a way that they feel would be beneficial to them.

**Random Response Bias.** Random response bias can occur when a respondent honestly does not know the answer to the question but answers anyway. This can happen when you ask a respondent to answer a question for which they would not reasonably know the answer. Respondents resort to guessing or speculating rather than reporting factual information. An example of this would be asking someone to indicate the motive of another individual, prompting a random response bias.

Another way this bias can manifest is when an individual has an opinion but hasn't considered their true feelings carefully. Like a central tendency bias, these individuals also tend to choose options toward the middle of the response scale. At times, people with this bias will choose the exact middle point (on an odd numbered response scale) simply because they don't want to think about the issue or don't really care. This bias can also manifest itself maliciously when an individual intentionally responds in a random fashion without actually reading the items. This can happen when there is an incentive involved and people simply want to complete the survey for the promised reward. If you suspect the possibility that a random response bias might occur, trigger items can be added to identify suspect response patterns. For example, adding an item that is the opposite of another item or reverse scoring an item can be used to identify potential problems. Note however, the practice of reverse scoring some items but not all the legitimate items is not advised as this can cause a response bias which can affect the analysis and interpretation of the results (Kulas et al., 2018).

## Chapter Summary

- It is easy to create a flawed survey. In fact, all surveys will be flawed to some degree. Having a flawed survey is more likely to happen when the design of the survey does not consider and attempt to alleviate potential problems.
- The most common mistake many researchers make when conducting survey research is administering the survey before it has been properly tested.
- Hiding flaws in a survey is unethical.
- Understanding how a survey may be flawed can help researchers create better instruments.
- Survey error is caused when researchers make mistakes designing and administering the survey, as well as when they misinterpret results.
- Response bias is caused by psychological influences that affect the way people respond to items on the survey.
- Survey error affects the validity of the results, meaning the survey does not provide a proper descriptive measure of what it was designed to measure. Survey error can also affect the generalizability of the results.
- Response bias adversely affects the accuracy (truthfulness) and reliability (consistency) of the results obtained. Bias influences the ability, or willingness, of participants to answer questions precisely or truthfully.
- Never underestimate the propensity of individuals to misunderstand what is being asked or otherwise provide information that is inaccurate to some degree. It's not a matter of you understanding the questions, but rather the respondent understanding.

## Discussion Questions

1. Explain how sampling and response rate affect generalizability.
2. Explain how scope error affects validity. Suggest ways in which this could be avoided.
3. For a survey topic you might wish to explore, identify various biases that would likely need to be addressed. What might you do to alleviate the potential problems?

# References

Kulas, J. T., Klahr, R., & Knights, L. (2018). Confound it! Social desirability and the “reverse-scoring” method effect. *European Journal of Psychological Assessment*.

# Item Writing

An important part of survey research involves the creation of good items. It's easy enough to write items, easier still to write poor items. Care needs to be taken to ensure the items align with the research purpose and questions. More importantly, once items have been constructed, researchers need to test the items to make sure the items are not flawed in any discernable manner. Creating good items is facilitated by following basic design principles.

## General Principles for Creating Items on a Questionnaire

There are two parts to any survey question, the item stem and the response options. The item stem poses the question or presents a statement. The response options provide a way for the respondents to answer the question or indicate the degree to which they agree with the statement. Depending on the questions, direct responses (i.e., open-ended or numerical inputs) can also be used rather than selecting options from a list. In addition, many survey software packages provide alternative ways for people to respond to a survey; however, while often creative, these can be hard to interpret and report. While there is not any particular order in which the following principles should be addressed, each needs to be considered when creating good survey items.

### Item Indispensability and Purpose

There should be a reason for including each and every item on a survey. Items must align with the study's purposes and objectives. As you write each item, consider how the response will help answer the research questions or support a specific research purpose. Ask yourself whether the item is needed or if additional items are required in order to fully understand the response (or use the results).

On some surveys, researchers will ask for information that is not needed, given the purposes of the study. Researchers might also ask for information they do not plan to use but for some reason feel might be interesting. For example, researchers often attempt to obtain demographic information they hadn't planned to use about the respondents. If these demographics do not serve a research purpose, you might consider removing the item from the survey to limit the length of the survey and avoid survey fatigue.

On the other hand, some research fails to ask enough questions. An item on the survey may provide broad impressions but fails to provide enough in-depth understanding to be useful. Two examples of this involve specificity issues and negative case analysis.

### Item Specificity Issues

Consider a basic survey item asking the degree to which individuals feel a particular intervention was effective. Obtaining an overall impression of effectiveness may satisfy the research needs; however, an individual may be generally satisfied with the intervention but more or less satisfied with particular aspects of the program. The level of specificity needs to be established. Abstract or complex topics may require multiple items each designed to capture nuanced aspects of the topic. This is especially the case when measuring constructs (see Affective Scale development).

## Negative Case Analysis

A negative case analysis can be used to better understand various responses. For example, a negative overall impression of satisfaction may require further investigation. For those individuals who felt generally satisfied with the intervention or program, no additional information may be needed; however, you may wish to know more from those who were less than satisfied. This would require branching and possibly an open-ended response option. You would need branching to avoid asking satisfied respondents unnecessary questions. If you don't already know all the possible things respondents might be dissatisfied with, you may need to use open-ended response options to ascertain which specific aspects of the intervention the respondent felt were unsatisfactory. There may be more than one problematic aspect, and some of the problems may be more egregious to the respondent than others—which may prompt you to ask participants to provide more information on those problems.

## Branching

Branching is a technique used to reduce the number of items presented to respondents or target individuals by only asking questions if the respondent belongs to a specific group. Group membership is often determined by key questions that can be used to modify the content or flow of the survey. For example, a respondent may be asked whether they have children living at home. The survey might use this information to present additional questions to individuals who answer affirmatively.

## Direct Sources

This principle involves making sure the questions you are asking can be answered by those you are surveying. You should never ask a question that requires the respondent to guess. This enables random response bias. For example, asking someone to speculate on the motive of others would not produce valid information. The information you obtain would represent the perception of the respondent (their guesses), not the actual motivations of those involved. It is unlikely that a reasonable purpose of any evaluation or research study would be determining speculative perceptions (guesses) individuals might have. It would be better to ask individuals about their own motives (asking those with a direct knowledge).

## Clarity and Precision: Audience Appropriate Wording

Surveys should be written for a specific audience. The vocabulary and structure of the item stem must be appropriate for the intended audience. To do this, you must understand those in the target population. When creating items, consider the target population's ability to read and how they will interpret the questions. Use common understandable language (i.e., natural language the target audience would easily understand). Pilot test each item with potential participants to verify the target audience would likely understand what is being asked. Just because you understand what you are asking doesn't mean they will. Be clear and precise when writing items stems. Avoid double negatives and keep items relatively short.

## Singular Purpose (Double-Barreled Items)

A common mistake many survey developers make is asking two related but separate questions in one item. We call this a double-barreled item. For example, a survey might ask, "How useful were the assignments and feedback?" These questions are tricky for respondents to answer because they don't know which aspect of the question to focus on (i.e., the usefulness of the assignments or the usefulness of the feedback). It is problematic for the researcher because the interpretation of the results becomes incredibly difficult, if not impossible. Double-barreled items should be avoided. They can usually be fixed by separating the item into two questions.

## Avoid Loaded or Leading Questions

A loaded question is one that has emotionally charged connotations that influence feelings and response patterns. A leading question is one that indirectly or unintentionally influences respondents to answer in a specific way. Loading a question with words that provoke emotional responses (positive or negative) will skew results. Using phrases that lead a respondent to make a connection or association when answering a question can also sway the way people respond.

The following examples show how an item stem statement might be loaded with emotional words and lead or influence individuals to respond a particular way. The first example loads the response by using the words "wrong," "condone," and "killing" when referring to abortion. This would likely skew the result against abortion. The second example, phrased as a women's rights issue using words like "rights," "access," and "choice," would likely skew results in the opposite direction. In both cases the wording would likely lead respondents to agree with the statement even though each represents opposite stances towards abortion.

> *Abortion is wrong because society should not condone the killing of innocent unborn children.*
>
> *Women have rights and should be allowed access to abortion services if they choose.*

## Item Stem: Response Scale Alignment

The item stem must align with and be appropriate for the response scale. In addition, respondents must be able to clearly understand each option and how to indicate their response. This can only be done by pilot testing the instrument.

## Pilot Testing

Before implementing a survey, the items should be tested. This can take various forms and phases, which may require several iterations. A think-aloud technique is commonly used to test a specific item. This involves asking a few individuals (preferably ones from the target population) to read aloud items on the survey and express their feeling and thought processes as they take the survey. If revisions are required, the process may need to be repeated. This would include determining if the item is clear, whether the wording might be improved, how the stem was interpreted, why the individual chose the response they selected, and if they felt the response options provided were adequate or whether additional options might be needed.

### Chapter Summary

- It is easy to create flawed items. Carefully considering various design principles will help improve the items you write.
- There are two parts to a survey item: the item stem and the response scale.
- There should be a clear purpose for including each item in the survey.
- Consider whether the items that were included are sufficient to answer the research questions and whether they are needed and adequate for the research purposes.
- Additional items, like a negative cases inquiry, may be needed to fully understand participant responses.
- Branching can be used to present specific items to targeted participants.
- Only ask participants to answer questions for which they would have a direct knowledge.
- The item stem should be carefully worded so the target audience would clearly understand its meaning.
- Avoid double-barreled item stems.
- Avoid loaded or leading item stems.
- Response scales should be appropriate for the item stem used.
- All items should be pilot tested, including a review by a representative of the survey's intended target audience.

## Discussion Questions

1. Explain how the pilot testing process might be used. What protocols and procedures might need to be implemented?
2. What tradeoffs need to be made to keep the survey reasonably short and still ask enough questions to answer the research questions?

# Section Contents



# Response Options

There are many different types of response scales; multiple-choice options are common, but Likert scales are used quite often as well. Each method used to obtain a response has limitations and strengths, as well as specific design principles that should be considered.

## Multiple-Choice Response

In this situation, the item stem is phrased as a question followed by a list of potential answers. Each response option must be mutually exclusive or clearly different from the others. There are a few specific design decisions you need to consider when using the multiple-choice response method.

### Example of a Multiple-Choice Response

Which of the following situations best describes your current relationship status?

- Married
- Widowed
- Divorced
- Separated
- Legal domestic partnership or civil union
- Single, but cohabitating
- Single, never married
- I am unwilling to share this information

### "Best Answer" or "Choose All That Apply"

Normally a respondent is expected to choose the answer that best represents their situation, beliefs, or condition. However, if the situation warrants, respondents may be asked to select all that apply to them personally. Only use the “all that apply” instruction if you have a good reason for doing so.

In the relationship status example, selecting the best answer is most appropriate. The options are mutually exclusive and expected to describe every relevant situation.

### Opting Out

Obviously, we want respondents to answer every item on the survey. Unfortunately, ethical considerations may require you to allow individuals to choose not to answer a specific question. In reality, you can’t force individuals to answer any question whether you provide an opt-out option or not. If an individual selects the opt-out option or simply chooses not to answer the item, the result is the same. Most research protocols involving the protection of human subjects will require you to allow respondents to opt out at any time. However, failing to answer a key question may equate to

refusing to take the survey, even if a respondent answers all the other questions. If the item was to be used as a disaggregation or branching variable, you should require respondents to answer the question (noting that respondents should still be allowed to opt out if they choose). If this is the case, you might not be able to use any of the respondent's data if they don't answer key questions.

Consider the relationship status item. This item most likely would be included in a survey to provide a basic description of respondent demographics. If this were the case, an individual's failure to respond may not be a problem. Providing respondents with an opt-out option for this item protects the individual's rights. However, if the results were intended to be disaggregated based on the relationship status, failure to answer the item would make the respondent's data unusable. This would be even more problematic if large numbers failed to answer or the response refusal had a systematic pattern. Similarly, if the information was to be used as a branching variable (so you could ask group-specific items), failure to capture this information would render this design feature inoperable.

## Number of Options (Option Specificity)

In most cases when using a multiple-choice response method, you want to provide a list of all reasonable responses, but not always. In the relationship status example, the number of possible responses will depend on the various mutually exclusive relationship situations possible. However, the degree to which specific details are required should be determined by the research purposes. You may only need to know if the person is single (never married) or something else. If this were the case, and you had no plans to use the more detailed information, you may only need two response options. In a multiple response scale, the researcher either attempts to make the list of options exhaustive, or they will need to allow "other" responses. However, If the list gets too long, there is the potential for fatigue and option order affect bias.

## Allowing "Other" Responses

When using the multiple-choice response method, the use of an option allowing respondents to write in another response has its advantages but also has limitations when it comes to reporting, interpreting, and drawing conclusions. The "other" option is used when the researcher is unsure whether or not they have an exhaustive list of possible answers. It should be noted that obtaining a comprehensive set of responses should have been obtained in the pilot testing phase of the instrument, which would reduce the need for allowing other responses.

To create a comprehensive list of possible answers, you might simply create a list using your own logic. For example, in the relationship status example it is very likely that you could logically determine each possible situation. However, this is not always the case. Consider a situation where you were looking for respondents' reasons for doing something. In the pilot testing phase, you might include an open-ended question asking potential respondents to state their reasons. Responses would then be coded and categorized. A comprehensive list of the most common reasons might then be obtained. This would be the basis for the list of answers you use in the final version of the survey. In this situation, you may still wish to allow for other responses. However, you would need to clean the data afterwards by reviewing each response not already listed. This is done by reading each additional response and deciding whether the response is novel (something you hadn't thought of), irrelevant/unrelated (the equivalent to a non-answer), or just a different way of saying what was already in the list of options (in which case the response should be recoded).

If the additional response is both new and meaningful, future versions of the survey should be updated to include this answer. However, the interpretation and reporting of the result you obtained with the incomplete set of response options would be difficult. If you report that one person provided an "other" response (a factual statement) you may be underreporting the importance or significance of the answer based on the frequency or prominence of the response. Had others been presented with the option, they may have also selected it as a response. This is an issue because we, as researchers, just don't know.

# Likert-Type Response Scales

A Likert response scale is most appropriate when the item stem is a statement. Likert response scales are typically used when designing a scale rather than a questionnaire but are also used in questionnaires. The Likert response scale is meant to capture the strength of an individual's agreement or belief. An adaptation of the Likert scale is often used to group individuals based on some numerical amount (e.g., income or age grouping). Response options in a Likert scale need to be ordinal in nature and mutually exclusive; multiple-choice response options need only be mutually exclusive. There are a few design decisions you need to consider when using this specific type of response scale.

### Examples of Items Using a Likert Response Scale

Police salaries are quite low.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree
- I don't know or have no opinion

About how much time (in hours) do you typically spend TEXTING each day?

- None (0, I don't text)
- Less than 1 hour (but not zero)
- 1 to 2 hours
- 3 to 4 hours
- 5 hours or more

## Ordinality and Mutual Exclusivity

This was mentioned previously but is important enough to explain further. A characteristic of ordinal data is that each point on the continuum represents a measurable increase in the amount or magnitude relative to the previous value on the scale and is some amount less than the next value on the scale. However, unlike interval-level data, the difference in the amount between each point on the scale is not neccesarily consistent (i.e., equidistant). For both ordinal and interval-level data, each point must be separate and distinct (mutually exclusive with no overlaps). In the texting example, if you provided the options "1 to 2" and "2 to 3," the options would not meet the mutually exclusive requirement because respondents with exactly two hours could select both answers. Likewise, the following set of response options intended to capture the amount of time spent on an activity does not meet the ordinality requirement because "multiple times a week" and "as little as possible" are ambiguous. Respondents may interpret the amounts represented differently than what was intended (e.g., "as little as possible" may actually be a lot, not a little; "multiple times a week" might be less than twice a day). Cognitive load is also increased as respondents struggle with trying to determine the amount of time each option represents.

### Example of Non-Ordinal Items in a Scale

- Multiple times a week
- Twice a day
- Once every day
- Every other day
- A few times a week
- As little as possible

## Option Specificity (Number of Response Options)

This is an issue for both multiple-choice responses and Likert response scales but for different reasons. With Likert response scales, the issue is one of specificity and, in some cases, ordinality. Typically, a Likert scale will have four to eight options. The specific number should be informed by the respondents' ability to make clear differentiations between options on the scale. Having too few options (e.g., "agree" or "disagree") limits the ability to interpret the data regarding the magnitude of agreement or disagreement. In addition to not being needed, having too many options (e.g., more than eight) makes it difficult to give meaningful labels to each response (i.e., labels that makes them distinct and clearly ordinal in nature). Respondents most likely won't be able to make meaningful differentiations between points on the scale, and less granularity would likely serve the research purposes.

### Examples of Option Specificity

Four options:

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Six options:

- Strongly Agree
- Agree
- Somewhat Agree
- Somewhat Disagree
- Disagree
- Strongly Disagree

## Direct Entry Option

One way to alleviate the specificity and ordinality problem, in situations where amounts are used to determine group membership, would be to have participants enter the amount directly or derive the amount based on some other data (e.g., determining one's age based on their birthday). Group cut points could then be established based on precise amounts rather than broad categories. Category cut points could be established in an a priori manner or based on natural divisions occurring in the data. If this approach is used, you will need to make it clear to respondents what units

are being used (e.g., days, hours, or minutes) as well as whether partial amounts can be entered (e.g., decimals or fractions). With online surveys, application software often has features that allow researchers to set restrictions for data entry to ensure amounts are within a valid range and formatted properly.

## Odds and Evens

Another often debated topic is the inclusion of a middle point. Having an odd number of options is sometimes used, which make it possible to have a middle point. Too often, however, it is used inappropriately. Having a middle point is acceptable when there is a true middle position. For example, when you are attempting to determine the amount of increase respondents believe occurred due to some phenomenon. If there was no increase and there was no decrease, a middle point representing no change is appropriate. Including the “neither agree nor disagree” option is thought by some to be appropriate in certain situations depending on the statement posed in the item stem; others feel it can lead to lazy responses and central tendency error when respondents don’t consider their choice carefully enough. Removing the middle point by having an even number of options forces respondents to decide whether they agree more than they disagree. It is never appropriate to place an opt-out option (e.g., “I prefer not to answer,” ”not applicable,” or “I don’t know”) in the middle of the response scale.

### Example of Middle Point Usage

Agreement or disagreement:

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

Increase or decrease:

- Greatly increased
- Increased
- Remained the same
- Decreased
- Greatly decreased

## Item Stems as Statements

Some practitioners recommend using questions (with multiple-choice responses) rather than statements and Likert scales. Both can be used effectively and there are advantages and disadvantages to using either. However, for a statement to work properly with a Likert scale, the statement needs to represent a clear position rather than a middle of the road (ambiguous or noncommittal) declaration. For example, you might ask individuals whether they thought police salaries were OK; however, there is little value in knowing whether respondents strongly agree that salaries were OK because it is difficult to interpret differences between agreeing they were OK and strongly agreeing they were OK. In both cases the respondent feels police salaries are just OK. It's like saying, "I believe my hair is brown" and "I strongly believe my hair is brown." The reason for using a statement in the item stem is to determine the strength of an individual’s feeling towards the idea or concept. In addition to making sure the statement presents a clear position, the statement should not be something that everyone might agree with in general (e.g., "It would be nice to provide adequate salaries for police.").

In addition, information about respondents' feelings might require additional information to determine what people think should be done. You many need two questions because knowing how individuals feel about police salaries may not align completely with what they feel should be done about it.

## Example of Question vs. Statement

Question:

What should be done about police salaries? Salaries should be

- reduced a lot
- reduced a little
- left as is
- increased a little
- increased a lot
- I don’t know or have no opinion

Statement:

Police salaries should be increased.

- Strongly agree
- Agree
- Disagree
- Strongly disagree
- I don’t know or have no opinion

## Chapter Summary

- It is easy to create flawed items. Carefully considering various design principles will help improve the items you write.
- Two major types of response scales include multiple-choice and Likert-type response scales. Ranking and rating, as well as direct input items, require a different method for collecting responses.
- Usually respondents must pick the single response that best describes their beliefs. Using a "select all that apply" instruction should be used with caution.
- Allowing participants to opt out may be required but can affect the validity of the survey results.
- Allowing other responses beyond those listed may be necessary but can introduce additional problems with interpretation and reporting.
- Option specificity (the number of response options in the scale) is an issue for both multiple-choice and Likert-type response scales.
- For multiple-choice response scales, the options need to be mutually exclusive and distinct.
- For Likert-type response scales, the labels used to identify each option on the scale need to be ordinal in nature and mutually exclusive (clearly different).
- A middle point should only be used when there is a true middle position participants might take.
- When using a Likert-type response scale, the statement used in the item stem must be a clear statement that is not ambiguous, ambivalent, or a generally accepted declaration.
- Carefully pilot test all items you create. Revise as needed, then test them again before administering the survey.

## Discussion Questions

1. What ethical issues need to be considered when writing items? How can these affect the quality of the survey?
2. In the item stem, when might it be best to use a statement rather than a question?
3. What difficulties or problems occur when allowing respondents to select "all that apply" for a specific item?

# Designing a Scale

Unlike a questionnaire, a scale is designed to estimate the strength of a personal characteristic or disposition held by an individual. The characteristics measured by a scale are called constructs. Scales are most often needed when measuring constructs in the affective domain. Measurement in the *cognitive domain* attempts to determine what people know, understand, or are able to do intellectually (e.g., recall, understand, analyze, evaluate, synthesize, problem solve, or think critically). Measurement in the *psychomotor domain* describes the quality of an individual's ability to perform physical tasks (e.g., athletic or musical ability). We do not need a self-report instrument to measure these attributes; other more direct measures are available and are more appropriate. However, measurement in the *affective domain* targets an individual's interests, perspective, beliefs, attitudes, or personality traits. Scales are needed to measure affect because the best (and possibly only) way to obtain data is to ask individuals to self-report. In addition, because affective constructs targeted by a scale are complex abstractions, multiple items are required to obtain an accurate measurement (each item on a scale is designed to capture a nuanced aspect of the construct). The related items on a scale are combined to provide a numerical indication of the strength of the construct for individual respondents.

Anderson and Bourke (2000) explain that affective characteristics being measured must (a) involve feeling, (b) describe a somewhat permanent trait rather than a temporary state, (c) be targeted toward something (e.g., feelings about specific people, situations, or entities), (d) have a positive or negative direction (e.g., like or dislike), and (e) vary in degree or magnitude for different individuals. For example, physical characteristics (e.g., height, hair color, or even appearance) do not involve feeling and can be measured without the need of a self-report instrument. Likewise, anything that can be simply categorized does not require measurement (e.g., age, gender, or nationality). These characteristics can be determined through simply reporting. Attitudes and perceptions, on the other hand, are personal attributes that often can only be determined through self-report. And while general feeling may be useful, specific feeling regarding specific targets are more helpful, especially when they represent typical or deeply held perceptions rather than temporary feelings (e.g., while you may have an opinion about politics in general, you would likely have contrasting and distinct opinions about various aspects of the political process). Finally, in order for affective characteristics to be measured, they must have both direction and intensity (i.e., the potential for individuals to have more or less positive or negative feelings about the topic).

### The Affective Domain

Affective constructs (and dispositions) represent inherent qualities of mind and character. A person's innate disposition can refer to one's predominant or typical mental and emotional temperament (e.g., a positive or happy disposition); however, a disposition can also refer to a person's typical state of mind, tendencies, and inclinations. These are attitudinal dispositions rather than innate personality traits. Attitudinal dispositions are often the target of scales used in survey research. Measuring these constructs is important because attitudes affect an individual's behavior (see Fishbein and Ajzen's [1975] theory of reasoned action). Attitudinal dispositions, although in the affective domain, also have a conative influence in that while our behaviors are influenced by our feelings, beliefs, and opinions, they are moderated by free will and agency (Davies, 2009). According to Ajzen's (1991) theory of planned behavior, people typically align their behavior with their personal belief systems after taking into account how they believe others would perceive their actions. They may also experience conflicting intentions, values, and beliefs (Davies, 2009). For example, an individual may be capable of performing specific skills, understanding important concepts, recognizing the value of the behavior, or having a personal desire to behave a certain way, but they still choose to act in ways we might not expect due to other deeply held beliefs or compulsions.

# Dimensionality

Some scales target unidimensional constructs, others are multidimensional. For example, when measuring students' attitude towards school, a scale would need to be multidimensional because students' interests and attitudes are based on a variety of unidimensional aspects. A student's attitude towards school might be influenced by the individual's perception of school being a safe place (physically and emotionally), a welcoming or fun place (socially), or a good place to learn (academically). The attitude may also be influenced by views of individual teachers or the student's personal interest in learning (career goals). It is very possible that a student may, for example, have a positive perception of school with regards to social aspects, but no interest or a negative attitude towards the academic opportunities provided. Each dimension would be measured separately (using a variety of items), then conceivably combined to get an overall measure of the student's attitude towards school.

# Creating Scale Items

Anderson and Bourke (2000) describe two methods for writing items for a scale—the mapping-sentence and domain-reference approaches.

## Mapping-Sentence Approach

When using this approach, the item developer pairs a situation with a behavior (or list of distinct actions). The respondent will either be asked to select the behavior that best describes the way they would react given the situation or indicate the likelihood they would behave in the manner specified given the situation.

Before writing items, the item developer must carefully identify the construct (defining its theoretical nature) and describe in detail how the construct might manifest itself as a behavior (operationalizing the various aspects or facets). When identifying the core of the construct, think about the kinds of beliefs and attitudes someone with this characteristic might have. When operationalizing the construct, consider what actions or behavior would be exhibited by someone who embodied this characteristic. If providing a list of possible behaviors, each potential response should represent an action likely to be taken by individuals possessing the characteristic at different intensities (e.g., high, moderate, or low). If the target is to be more specific (rather than a general measure of the construct), the situation can be contextualized.

### Mapping-Sentence Example

Suppose you were attempting to measure the construct of self-confidence (in general). At its core, confidence requires a high level of self-efficacy, a belief in one's ability or competency. In operational terms, someone with self-confidence may try new things, attempt difficult endeavors, or be willing to take risks. They might also express feelings of success. This understanding could help you develop items to place in the scale. Pairing situations with behaviors we believed would be taken by someone with self-confidence allows us to get an indication of the strength and direction of the characteristic for individuals. While we might be interested in obtaining a general measure of confidence, contextualizing the situations allows us to get targeted measure of an individual's confidence in specific situations (i.e., ability to compete in a specific sport or succeed in a specific endeavor).

Possible situations:

- starting a new task
- trying a difficult maneuver
- giving an opinion
- learning a new skill

Anticipated behavior:

- begin without hesitation
- wait until encouraged
- think carefully whether to proceed
- let someone else try first
- believe they will be successful

Possible items:

- When learning a difficult maneuver for the first time, I immediately begin without hesitation.
- When learning a new skill, I am confident I will succeed.

## Domain-Reference Approach

This approach also begins with an examination of the construct's critical features. However, particular consideration is given to the target, direction, and intensity of the characteristic. Statements are created by specifying a target, using a directional form of the "to be" verb (positive or negative), and then adding an adjective denoting intensity. Once the basic structure of the statement is determined, it can be transformed to fix grammatical issues. The statements produced should have a direct alignment between conceptual and operational definitions of the construct.

### Domain-Reference Example

Suppose you were attempting use this approach to write items that could be used to measure the construct of self-confidence. Possible targets are listed. A verb is selected, followed by an adjective that describes a possible reaction. The statement should describe a precise behavior and a strong emotion or position.

Target: learning a new skill, participating in a debate, expressing my opinion

Verb: I am, is

Adjective: exciting, thrilling, eager, happy

Possible revised statements:

- I am eager to try difficult maneuvers for the first time.
- I am happy to express my opinion to others.
- I usually express my opinions quite well when debating.
- I am very good at learning new skills.

## Testing and Revision

As with all survey questions, items must be carefully tested to make sure they are meaningful and clear to potential participants. Pilot testing the items using think aloud cognitive interviews can be quite useful for this purpose. Multiple revisions may be needed. Some additional guidelines include the following:

- Write multiple items that cover all aspects of the topic and likely targets. Test each item and use those that work best.
- Keep the language clear, simple, direct, and concise (see writing clear items).
- Each statement should contain only one thought or situation (see double barreled items).
- Avoid statements of fact.
- Avoid general statements everyone or no one would likely endorse. Write statements that represent strong, clear positions rather than a middle of the road (ambiguous or noncommittal) declaration.
- Write in the present tense using an active voice.
- Avoid the use of vague modifiers (e.g., usually, sometimes).

## Chapter Summary

- A precise, agreed-on theoretical definition for the construct being measured is required before writing items. This includes both conceptual and operational definitions.
- Conceptual definitions describe the abstract meaning of the construct and the core beliefs or motivations foundational to the construct.
- Operational definitions describe how those with the characteristic would likely behave. Manifestation of anticipated behaviors provides inferential evidence that the individual possesses the characteristic.
- A close alignment between conceptual and theoretical definitions should exist and be evident in each item.
- Scales are most often needed when measuring personal characteristics in the affective domain.
- Characteristics in the affective domain must (a) involve feeling, (b) target stable traits rather than temporary states, (c) be specifically targeted (d) have direction (positive or negative, absence to increasing positive), and (e) vary in intensity by degrees.
- Two approaches to writing items include the sentence-mapping and domain-reference approaches.
- Multiple items for each construct should be created and tested prior to pilot testing and administering the instrument.
- Statistical procedures can be used to validate the instrument, including exploratory and confirmatory factor analysis or structural equation modeling.
- Constructs may be unidimensional or multidimensional.

# Pre Data Analysis Activities

Once you obtain survey results, the fun begins. Note however that analyzing and reporting survey data can be challenging in that it can be a time-consuming process. If results are not processed, analyzed, and reported properly, the results may be misleading and possibly inaccurate. There are several issues to consider and a few things to do before you get into the actual data analysis, interpretation, and reporting.

## Cleaning and Organizing the Data

Before data analysis can take place, the information you obtained must be cleaned and organized. Cleaning data refers to the process of removing irrelevant data (as in the case where online surveys add variables to facilitate the survey's function), possibly de-identifying the responses (as required by IRB protocols), or coding open responses (see allowing "other" responses). Cleaning data is needed prior to examining response patterns and identifying incomplete surveys. The data may also require reorganizing (e.g., collapsing categories or performing summary calculations). For example, you may wish to report the percentage of those who "agree" or "strongly agree" with a statement in comparison to those who "disagree" or "strongly disagree." The precise cleaning needs will be determined by the purposes and reporting needs of the survey.

## Systematic Nonresponse Review

Once the data is cleaned, a review of the data is needed to ensure no systematic survey nonresponse patterns have occurred (see Response Rate Issues). Even when random sampling is used, you may have a problem if response refusal patterns indicate that an important group of potential respondents failed to complete the survey. Key indicators may include age, gender, or race patterns that do not match the expected demographics. However, other response refusal problems may cause a lack of generalizability—for example, location overrepresentation (e.g., urban vs. rural) or group underrepresentation (e.g., freshman vs. seniors). The degree to which this will be a problem depends on the degree to which the sample provides a reasonable representation of the population. Without pertinent information about respondents, you will not be able to conduct this investigation. Therefore, it is important to identify key factors and plan to obtain these details prior to administering the survey. If an issue is identified, you may need to get additional responses from underrepresented individuals to solve the problem.

## Determining the Response Rate

One of the first things you should calculate once a survey has been administered and the results obtained is the response rate. The calculation is simple. The response rate is an indication of the number of invited participants who complete the survey. This is reported as a percentage but should always be accompanied by the size of the sample. A 66% response rate obtained from a sample comprised of three individuals wouldn't provide very compelling evidence. However, a response rate of this magnitude obtained from a much larger sample would be much more impressive. You should also indicate how the sample was obtained when reporting the sample size (the population size as well if it is known). The sampling procedures will have been detailed in the methods section, but you should briefly report them (as a reminder) when presenting results (e.g., a random sample of three individuals).

$$Response\ Rate = \frac{Number\ of\ Usable\ Surveys\ Returned}{Number\ of\ Invites}$$

## Completed Survey Decisions

The only consideration (possible controversy) surrounding the response rate calculation is deciding what it means to complete the survey and whether information obtained from partially completed surveys might be usable. In order to make this determination, a careful examination of the survey results is needed. While there are no hard and fast rules, there are some principles that may help you make this determination. In situations where a returned survey is unusable, it should not be included in the response rate calculation. This will also affect the margin of error calculation.

**Requisite Data.** A guiding principle for making an inclusion decision is to determine what information is essential in order to answer the research question. Sometimes a partially completed survey can be used to answer some of the research questions. Other times, missing a single item on the survey will render the information unusable. For example, if one section of the survey was completed and not another, part of the information provided might be usable. However, the information provided in a partially completed survey would likely need to be excluded if, for example, the respondent only completed the demographics section but nothing else, or they failed to provide vital grouping information required to disaggregate the data and answer the research question.

**Accurate Data.** Another factor that should be considered regarding the completeness of the survey is that of accuracy. Unfortunately, there are times when participants are not completely honest in the way they answer questions on a survey. You may not know how accurate the information provided will be, but you can get indicators that the information is inaccurate. For example, suppose that while pilot testing the survey you determine that a survey typically takes 10 to 15 minutes to read, reflect, and answer all the questions. Then suppose a participant completed the entire survey in only two minutes. You might suspect the results to be inaccurate and quite possibly unusable.

Inaccurate (unusable) data might also be identified by examining a participant's response pattern. For example, the results might be suspect if a respondent provided the same response for every question, even though it would be extremely unlikely that an honest respondent would actually answer in that way. Random response bias like this is more likely when incentives are provided to individuals for taking the survey. If trigger items were added to identify suspect response patterns, these would need to be examined (see random response bias).

## Acceptable Response Rates

Knowing the response rate is an important piece of information. Opinions about what constitutes an acceptable response rate are debatable. Ultimately, you need to feel confident that the information you obtained will adequately serve the purposes of the survey and answer the research questions. For some research purposes, only near-perfect response rates with no perceivable systematic response refusal patterns would be acceptable. In other situations, a much smaller response rate might suffice. Response rates for published social science research can range from 25%–75%, and a 30% response rate is typical (Baruch & Holtom, 2008). It is up to the researcher to persuasively argue that the response rate obtained is sufficient.

# Estimating the Margin of Error

For some surveys, you may need to calculate the margin of error. You can calculate a margin of error for each of the values obtained from the survey. The margin of error is an estimate of the amount of error we might expect for each outcome. In practice, the margin of error is a confidence interval. Any statistic we obtain from a survey is an estimate that includes some amount of error. We don't actually know what the real (true) value is, but we can be somewhat confident that the true value will fall within a specific range based on an estimate of the standard error (SE) and a specified confidence level (z).

# Margin of Error Calculation (proportion)

In order to calculate the margin of error for a result represented by a proportion, we need three values: the confidence level, the sample size (i.e., number of completed surveys with usable data), and the sample proportion. If the population size is known (i.e., you have a finite population), the formula can be adjusted to account for any error that might occur from using a sample instead of taking a census. The modified formula (using the Finite Population Correction, or FPC) assumes you know the population size.

$$Standard\ Error = \sqrt{\frac{p(1-p)}{n}}$$

$$Margin\ of\ Error = Z\sqrt{\frac{p(1-p)}{n}}$$

$$Margin\ of\ Error_{(fpc)} = Z\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

- p = the sample proportion
- n = sample size (number of usable surveys)
- N = population size
- Z = z-value representing the desired confidence level

  2.576 for 99% level of confidence

  1.96 for 95% level of confidence

  1.645 for 90% level of confidence

## Margin of Error Example for a Proportion

Continuing with the counseling services example from the beginning of this book, let's say you obtain a result where 84% of respondents selected a specific option on one item (p=.84). This means 16% selected a different option (1-p or .16). Suppose we decided to use a 95% confidence level, which would make Z=1.96. Given a sample size of n=180, that would make the margin of error equal to 2.73. Given this margin of error, we can say the result is assumed accurate within plus or minus 2.73 percentage points with a 95% confidence level. However, adjusting for the fact that this is a finite population (N=5000), an adjusted estimate suggests the margin of error might actually be 2.68. These estimates are quite close and both round to 2.7 percentage points. This means the statistic obtained might reasonably be anywhere between 82.3 and 86.7 percent (84 ± 2.7%).

$$Margin\ of\ Error = 1.96\sqrt{\frac{.84(1-.84)}{180}} = .0273\ or\ \pm 2.73$$

$$Margin\ of\ Error_{(fpc)} = .0273\sqrt{\frac{5000-180}{5000-1}} = .0268\ or\ \pm 2.68$$

# Margin of Error Calculation (Mean Values)

In order to calculate the margin of error for a result represented by a mean, we need three values: the confidence level, the sample size (i.e., number of completed surveys with usable data), and the standard deviation of the sample mean. If the population size is known (i.e., you have a finite population). The formula can be modified using the FPC.

$$Standard\ Error = \frac{\sigma}{\sqrt{n}}$$

$$Margin\ of\ Error = Z\frac{\sigma}{\sqrt{n}}$$

$$Margin\ of\ Error_{(fpc)} = Z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- σ = standard deviation of the sample mean
- n = sample size (number of usable surveys)
- N = population size
- Z = z-value representing the desired confidence level

## Example

When the survey result is a mean rather than a proportion, the standard error calculation uses the standard deviation of the sample mean. Suppose you asked people how often (in days or weeks) they experience feelings of depression, and you determine that the average response was two days with a standard deviation of 2.8 (σ = 2.8). Using a 95% confidence level and given a sample size of n=180, the margin of error would be 0.409. Given this margin of error, we can say the result is assumed accurate within plus or minus 0.409 days at a 95% confidence level. However, adjusting for the fact that this is a finite population (N=5,000), an adjusted estimate suggests the margin of error might actually be 0.402. These estimates again are quite close, around 0.4 days. This means the statistic obtained might reasonably be anywhere between 1.6 and 2.4 days each week (2 ± 0.4).

$$Margin\ of\ Error = 1.96\frac{2.8}{\sqrt{180}} = .409$$

$$Margin\ of\ Error_{(fpc)} = .409\sqrt{\frac{5000-180}{N-1}} = .402$$

## Chapter Summary

- Data cleaning is required prior to starting the data analysis process.
- If results are not processed, analyzed, and reported properly, the results may be misleading and possibly inaccurate.
- Data should be reviewed to identify unusable surveys and explore the possibility of any systematic response refusal pattern.
- Response rates, along with sample size and sampling methods, should be reported.
- An adequate response rate is needed to obtain a representative sample.
- Each statistic obtained from a survey is only an estimate of the true population parameter.
- A margin of error calculation can be used to provide a confidence interval for each sample statistic.

### Discussion Questions

1. What impact would you expect if you found a systematic response refusal pattern had occurred? What step might you consider taking to alleviate the problem?
2. How does the response rate affect the sampling process?

## References

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human relations*, *61*(8), 1139–1160.

# Data Analysis Basics

There are a variety of ways data can be analyzed. Choosing appropriate methods is important. Presenting (displaying) and reporting (interpreting) data properly is also essential.

## Descriptive Statistics

Descriptive statistics are used to summarize information obtained from the sample without making any direct claims about the population. Descriptive statistics are used to present the sample data in more meaningful ways, which helps us understand and interpret the data later. While descriptive statistics are meant to summarize and present survey results, you may want to point out interesting aspects or patterns in the findings, but you don't make explicit inferences or generalizations about the population yet. Common visualizations of survey results include bar charts, frequency distributions, or pie charts. Tables can also be useful for displaying descriptive data.

## Inferential Statistics

Inferential statistics are used to draw conclusions (inferences or generalizations) about the population from which a sample was drawn. Statistical techniques will use confidence intervals (margins of error), regressions (predictions), or hypothesis testing (involving statistical and practical significance) to estimate something about the population based on the sample.

Statistical significance and practical significance are determined to provide evidence that the result has some importance. Statistical significance refers to the probability that observations in the sample may have occurred due to chance. Given a large enough sample, despite seemingly insubstantial results, one might still find a satisfactory level of statistical significance. Practical significance, on the other hand, looks at whether the magnitude of the observation is large enough to be considered substantial. For example, when considering the difference between the mean of two groups, you might find that a difference of 1% is statistically significant (e.g., it has only a 5% chance of occurring due to chance), but you realize that the magnitude of this difference has no practical significance (i.e., the difference is not really that different in practical terms).

## Types of Data

Prior to conducting your data analysis, you need to make sure you understand the type of data you have so you can select appropriate statistical methods. For certain types of data, it is inappropriate to use some statistical analyses.

There are four basic types of data, although many statistical programs combine interval and ratio data (calling it scale data) because the statistical methods used with these types of data tend to be the same.

### Nominal Data

Nominal data might best be described as categorical. These data are the most basic type of information you might collect in a survey. Rules are used to specify membership in a category. Frequency (group size, counting) and

proportional information (percentages) are used to report these types of data. These are also commonly used to disaggregate data when comparing groups. However, when making group comparisons, group membership rules should make it so that groups are mutually exclusive (i.e., no individual is a member of both groups being compared).

## Ordinal Data

These data have some sense of order, but the intervals between points on these types of scales are not equidistance. For example, placement results or preference (i.e., first, second, and third) have an order, but differences between various points on the scale are not consistent (first and second choices may be close, but both might be far more preferred than anything that comes next). Computing the mean and standard deviation for ordinal data is discouraged and, in most cases, inappropriate (although some researchers regularly compute averages for results obtained from Likert scales); frequencies (mode) and proportions (percentages) are best used when describing results based on this type of data along with ranking results. When making inferences, some nonparametric statistical procedures might also be appropriate.

## Scale Data (Interval and Ratio)

Scale data have all the properties of nominal and ordinal data but also have the characteristic of equal intervals; in the case of ratio-level data, they have a true zero point. This means the distance between each point on the numeric scale being used is the same regardless of where on the scale you look. For ratio-level data, this also means that comparisons can be made about differences in magnitude (e.g., twice as much). It is appropriate to calculate the mean and standard deviation of scale-level data. You can add and subtract interval-level data, but you can also multiply and divide ratio-level data. With scale data, in addition to means and standard deviations, inferential statistics can be used—including *t*-tests, correlations, and regression analysis.

**Table 1**

*Types of Data and Their Characteristics*

| Type and Characteristic | Typical Applications | Scale Characteristics Possessed | | | |
|---|---|---|---|---|---|
| | | Identity | Order | Distance | Origin |
| **Nominal** — identification or classification | Gender<br>School number<br>Geographical location | ⊗ | | | |
| **Ordinal** — specifies order or rank | Brand preference<br>Placement<br>Agreement (Likert scales) | ⊗ | ⊗ | | |

| | | | | | |
|---|---|---|---|---|---|
| **Interval** — specifies order based on equidistant intervals (implies equal increments of measurement) | IQ, test scores<br>Degree in F° and C°<br>Time of day | ⊗ | ⊗ | ⊗ | |
| **Ratio** — interval data with a zero point denoting an absence of the characteristic being measured. | # correct, Units sold<br>Distance, Time (amount)<br>Height, Weight, Age<br>Degrees in K° | ⊗ | ⊗ | ⊗ | ⊗ |

# Data Visualization

How you present results is important. Primarily used with descriptive statistics, tables, graphs, and charts summarize information in a readable format. These presentation methods not only organize large amounts of information, but they can also help focus readers' attention on patterns and important findings. They are often the basis from which inferential statistics are calculated. While this course does not elaborate on the data visualization theories and practices, several resources exist to help develop data visualization skills (see references for some examples).

## Chapter Summary

- Descriptive statistics are used to summarize survey results.
- Inferential statistics provide evidence used to support conclusions (inferences or generalizations).
- Data obtained from a survey will fall into various data types (nominal, ordinal, interval, or ratio).
- The appropriateness of the statistical analysis used is determined by the characteristics of the data (i.e., type of data).
- In survey research, perhaps the most controversial statistical issue pertains to whether data obtained from Likert scales can be used as interval-level data (i.e., assigning numbers to responses and averaging the findings).
- Data visualization theory and practices are extremely important for presenting descriptive statistics effectively.

# References

Evergreen, S. D. H. (2018). *Presenting data effectively*, 2nd Edition. Sage Publishing.

Evergreen, S. D. H. (2019). *Effective data visualization: The right chart for the right data*, 2nd Edition. Sage Publishing.

Data Visualization Presentation: Choosing Charts

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. Wiley Publishing.

# Data Analysis Basic Example

Perhaps the most common descriptive statistic obtained from a survey is a simple frequency count followed by a proportion (i.e., percentage of the valid responses). Specific frequency results are sometimes reported simply as a textual description (e.g., "40% of respondents indicated satisfaction"); however, results such as these are more commonly presented in tables or charts depicting the number of individuals who select each option on an item.

## Frequencies Example (Tables)

Most statistical packages will provide basic frequency counts and percentages. In this example, respondents were asked to indicate the highest academic degree they had obtained. When there is a lot of data to report, a table is an efficient way to present results.

| Highest Academic Degree Obtained | | | | |
|---|---|---|---|---|
| Response | Frequency | Percent | Valid Percent | Cumulative Percent |
| Highschool or equivalent | 86 | 17.2 | 16.8 | 17.4 |
| 2yr college (associates) | 62 | 12.4 | 12.5 | 29.9 |
| 4yr college (bachelors) | 208 | 41.5 | 42 | 71.9 |
| Masters or equivalent | 108 | 21.6 | 21.8 | 93.7 |
| Doctorate or equivalent | 31 | 6.2 | 6.3 | 100 |
| Total Valid | 495 | 98.8 | 100 | |
| Missing | 6 | 1.2 | | |
| Total | 501 | 100 | | |

You will note that six individuals failed to respond to this survey item; this means there was a 1.2% response refusal rate for this item. You will need to clearly indicate whether the proportions being reported represent the valid percentages (i.e., only those who responded to the item) or total percentages (i.e., all who returned the survey, including those who did not answer this question).

It may have occurred to you that in this example, participants were not offered a way to indicate they had not earned any academic degrees. You may also inappropriately assume that the six people who did not respond did so because they had not earned any academic degrees but had no way to indicate their situation. While possible, this assumption cannot be substantiated. It may be that some nonrespondents simply overlooked the question. Obviously, it would be better to have anticipated this possibility and included an option that allowed the respondents to indicate they had not earned any academic degrees.

In this example, cumulative percentages can be useful because the data is somewhat ordinal in nature. Note that it would be inappropriate to average the results. Calculating the mean is only appropriate with interval-level data. It would however be appropriate to point out important trends involving the mode.  For example, 41.5% of the respondents had earned a bachelor's degree and 69.3% of the respondent reported earned a bachelor's or graduate degree. Because this is a substantial number, it may be worth pointing out.

## Collapsing Categories Example

Depending on the research questions, it may be appropriate to collapse categories in order to better understand the data. Suppose you really only wanted to know how many people had earned a college degree of some kind. In this case, you may wish to reorganize the data in the table to present undergraduate and graduate degrees together. It might be important to point out that 81.6% of the respondents had earned a college degree of some kind. You may also need to collapse these categories so you can identify groups for disaggregation purposes.

| Highest Academic Degree Obtained | | | | |
|---|---|---|---|---|
| Response | Frequency | Percent | Valid Percent | Cumulative Percent |
| Highschool | 86 | 17.2 | 17.4 | 17.4 |
| Under Graduate | 270 | 53.9 | 54.5 | 71.9 |
| Graduate | 139 | 27.7 | 28.1 | 100 |
| Total Valid | 495 | 98.8 | 100 | |
| Missing | 6 | 1.2 | | |
| Total | 501 | 100 | | |

## Likert Scale Example

It is quite common for surveys to use a Likert scale to record responses. Because these data are generally considered to be ordinal in nature, it is appropriate to report frequencies rather than averages. Assigning point values for each response and calculating the mean (central tendency) is sometimes done, but doing this would not provide sufficient detail needed to fully understand the response patterns. Note also that, as with the previous example, it is sometimes appropriate to collapse categories when using a Likert scale. For example, "disagree" and "strongly disagree," "somewhat disagree" and "somewhat agree," and "agree" and "strongly agree" might be collapsed from six categories to three.

| Willingness to Encourage Child to Persue a Career in Teaching | | |
|---|---|---|
| Response | Frequency | Valid Percent |
| strongly disagree | 43 | 8.7 |
| disagree | 95 | 19.3 |
| somwhat disagree | 63 | 12.8 |
| somewhat agree | 117 | 23.7 |
| agree | 135 | 27.4 |
| strongly agree | 40 | 8.1 |
| Total | 493 | 100 |

| Willingness to Encourage Child to Persue a Career in Teaching | | |
|---|---|---|
| Response | Frequency | Valid Percent |
| No | 138 | 28 |
| Maybe | 180 | 36.5 |
| Yes | 175 | 35.5 |
| Total | 493 | 100 |

## Crosstab (Pivot Table) Example

Survey research is often table heavy. As a result, we often want to combine tables, especially when we want to disaggregate our data based on some grouping variable. Using a crosstab (sometimes called a contingency table or pivot table) we can combine tables. This allows us to present summaries of our data in a more efficient manner (reducing the number of tables needed). The challenge is to make sure results are well organized and clearly presented.

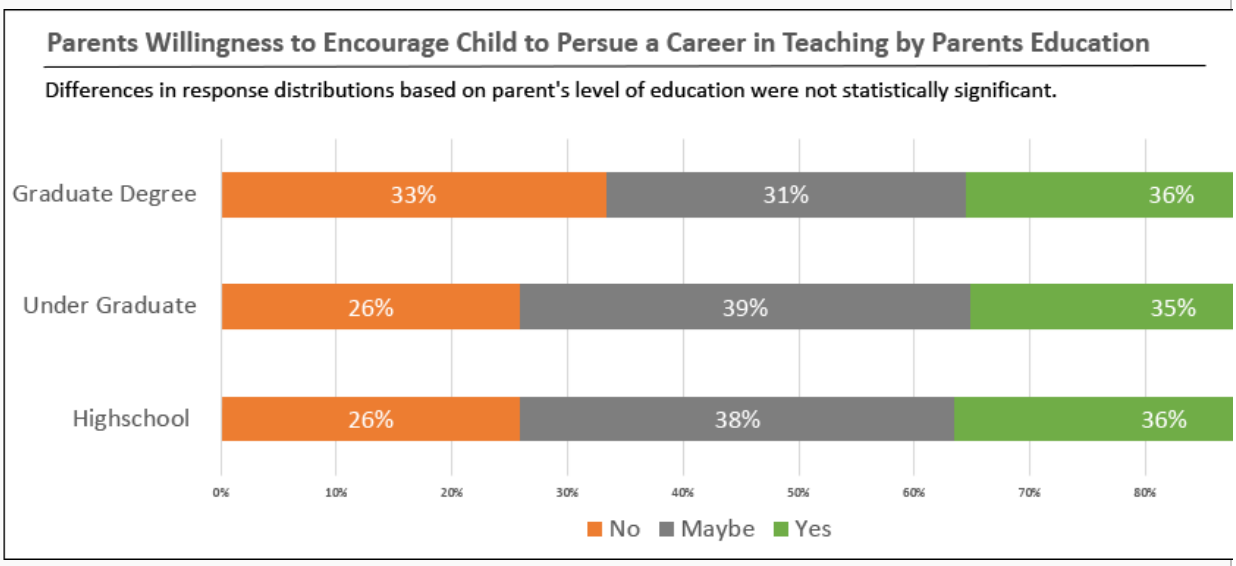
| Willingness to Encourage Child to Persue a Career in Teaching by Parents Education | | | | |
|---|---|---|---|---|
| | Indicated Willingness | | | |
| Parents Education Level | No<br>count(% within) | Maybe<br>count(% within) | Yes<br>count(% within) | Total in Group |
| Highschool | 22 (25.9) | 32 (37.6) | 31 (36.5) | 86 |
| Under Graduate | 70 (25.9) | 105 (38.9) | 95 (35.2) | 270 |
| Graduate | 46 (33.3) | 43 (31.2) | 49 (35.5) | 139 |
| Total | 138 (28.0) | 180 (36.5) | 175 (35.5) | 493 |

**Inferential Statistics**

The data presented above is descriptive in nature. It presents a summary of what the respondents reported, disaggregated by the respondents' education level. However, while we see some differences in the response distribution by group, we would need to use inferential statistics to determine whether these results are statistically significant (i.e., whether the difference could be attributed to chance). The appropriate statistical analysis to use will depend on the type of data obtained. Given that these data represent proportions, a chi-squared test of independence would provide evidence of whether any difference in the distribution of responses for each group was statistically significant. Note that if the result was found to be statistically significant, the practical significance should also be calculated and reported.

## Data Visualization Example

While tables can be a very effective way to present large amounts of data, often results can be presented more effectively using charts and graphs. Using proper data visualization techniques can enhance the presentation. Choosing the right chart type, as well as paying attention to colors, fonts, and layout, is important.



# Data Analysis Pre-Post Example

A potential reason for conducting survey research is to measure change. Quite often researchers want to be able to determine whether any change can be attributed to a specific intervention, but they might also simply be interested in identifying a change in general (not attributed to any specific causal relation).

Measuring change usually requires that we have two measures of the variable in question for each of the individuals responding to the survey. Change is calculated by determining the difference between the paired values. This typically means administering two surveys (i.e., pre and post surveys each providing a measure of the variable). It is best practice to use the same individuals for the pre and post surveys; however, in some cases, different groups of individuals might be surveyed if the samples are randomly selected representative samples of the population. Change, in this case, can be determined by comparing the average difference obtained from each survey. In some cases, a single survey (posttest only) might attempt to determine change by asking respondents to report the amount of change they experienced or by asking respondents to report pre and post levels of the variable in question.

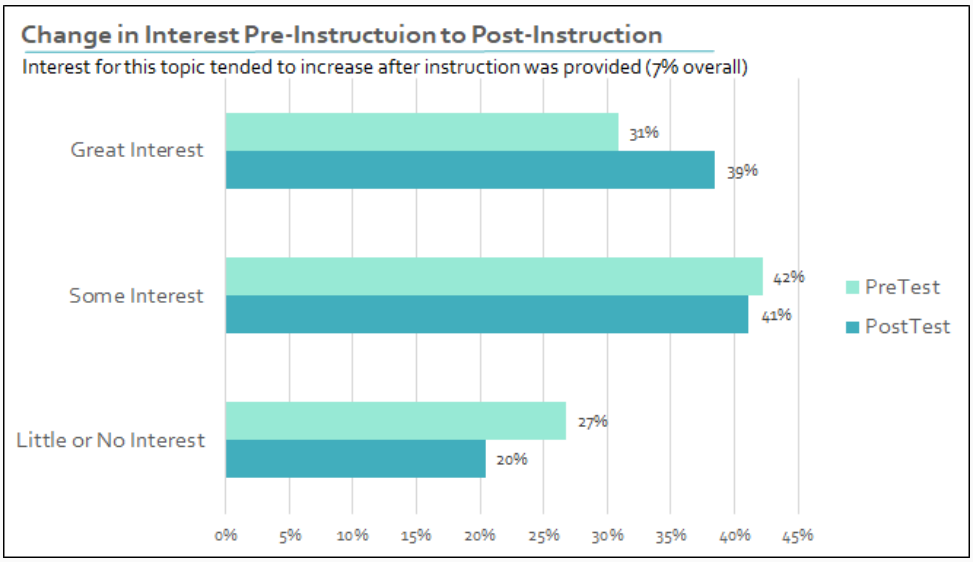
## Reporting Amount of Change

When data is interval or ratio level (scale data) the mean differences can be reported. However, in many cases, the data obtained in a survey is best described as ordinal data and should be reported as proportions or frequencies.

Consider the situation where you ask respondents about their level of interest in a topic (pre and post). You might speculate (or hope) that after receiving instruction, those involved might become more interested in the topic. In this case, the data you obtain is not interval-level data because the data points are not equidistant. We can still provide descriptive information about the change that occurred, as well as inferential evidence regarding the statistical significance and magnitude of the change.

Reporting average percentages by response categories, we see that in general there was a 7% increase in interest post-instruction. Using inferential statistics, you would be able to determine the statistical significance (probability that this result was due to chance) and the practical significance (whether the magnitude of the change was substantial). A chi-square goodness of fit test could be used for a pre-post comparison using the posttest results as observed values and the pretest results as expected values. Remember that it is best practice to also report the sample size and response rates.

| Interest Level Change Pre-Instruction to Post-Instruction (n=1469) | | |
|---|---|---|
| Interest Level | Pretest | Posttest |
| Little or No Interest | 392 (20%) | 300 (27%) |
| Some Interest | 623 (41%) | 604 (42%) |
| Great Interest | 454 (39%) | 565 (31%) |



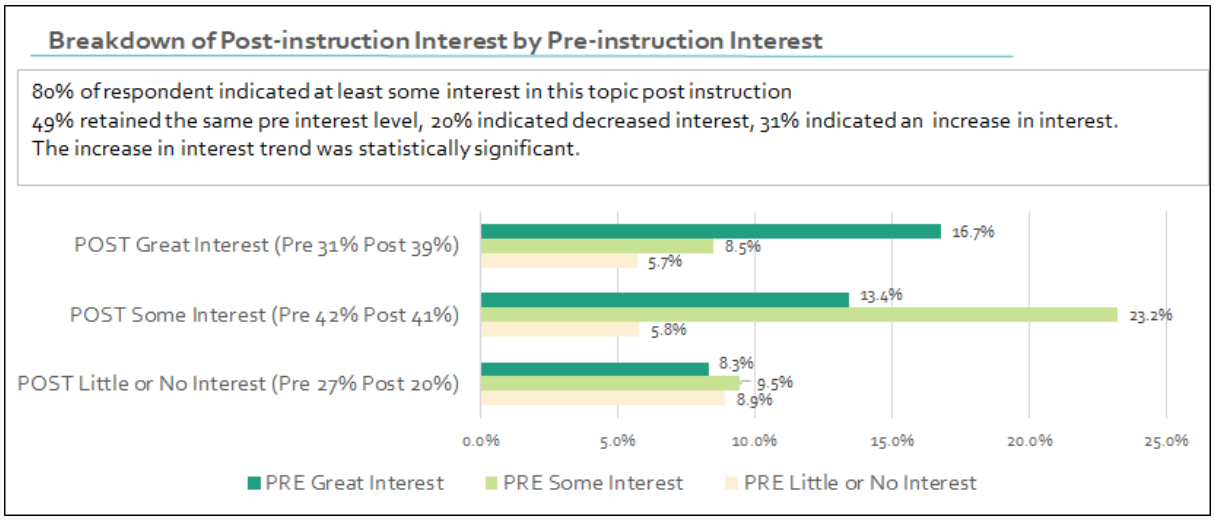
## Reporting Breakdown of Change

In this example, reporting overall change in interest can be a bit misleading. You might erroneously conclude that the change in interest was all positive. This is not the case. While the aggregate change has a positive trend, many individuals actually reported a decrease in interest after instruction. This change is not represented in the previous analysis. A more complete analysis would report the breakdown of the change, which includes the magnitude and direction.

Based on this analysis, while there was an overall increase in interest, 31% of respondents reported an increase in interest, and 20% of the respondents reported a decrease in interest. Nonparametric inferential statistics (e.g., a McNemar-Bowker test of symmetry) could be used to determine the statistical significance of the result.

Interest Level Breakdown of Change Pre-Instruction to Post-Instruction

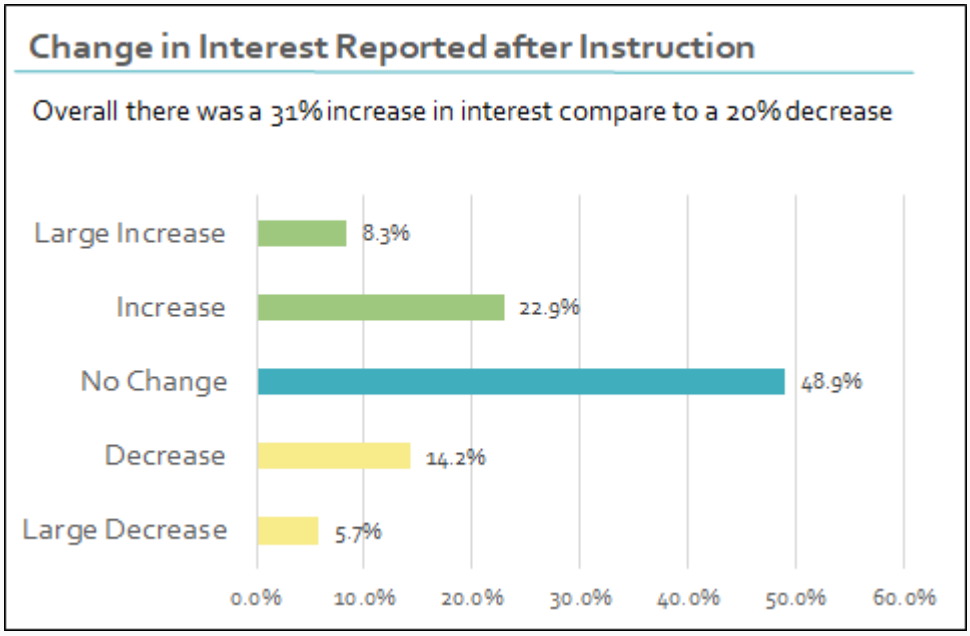
| | | Post Instruction | | | |
|---|---|---|---|---|---|
| | Interest Level | Little or No Interest | Some Interest | Great Interest | Overall Pre |
| Pre Instruction | Little or No Interest | 9% | 9% | 8% | 27% |
| | Some Interest | 6% | 23% | 13% | 42% |
| | Great Interest | 6% | 8% | 17% | 31% |
| | Overall Post | 20% | 41% | 39% | |



## Posttest Only Example

While this result may best be obtained using pre and posttest surveys, at times this is not possible. You may only be able to ask respondents (in a posttest-only survey) to indicate any change in interest they believe occurred after receiving instruction.

| Change in Interest Reported after Instruction | | |
|---|---|---|
| | n | % |
| Large Decrease | 84 | 5.7 |
| Decrease | 209 | 14.2 |
| No Change | 718 | 48.9 |
| Increase | 336 | 22.9 |
| Large Increase | 122 | 8.3 |
| Total | 1469 | 100 |




## Discussion Questions

1. What are the benefits and limitations of measuring change using data from match pairs, representative samples, or posttest-only, single-sample data?