Chapter 11: Sampling and Generalizability

Psychological science is a science of people, by people, for people.

In every psychological research study, we have a group of participants who are recruited to voluntarily participate in the research project. These participants are our **sample,** and they are drawn from a larger **population**. This is part of our statistical model as we take our descriptive statistics about the mean performance and variance from our sample and try to evaluate how accurate these numbers are with respect to the population. For statistics and our inferences about difference, the number of participants is a critical aspect of this calculation with more participants, or larger ‘**n**’, always being better.

In drawing inferences from our data, it is also necessary to consider where this sample came from, how it was recruited and how it might relate to or differ from the broader population. When we draw a conclusion from our data, who do we expect (hope) the conclusions will apply to? The answer to that question will vary across subdisciplines within psychology. Some areas are ambitious, such as cognitive psychology, which hopes to draw inferences that apply to all humans about memory, perception, or other basic human cognitive processes. Some areas are much more specific, such as developmental psychology, which may aim to draw inferences about behavior in a very specific age range (e.g., 6-month old infants, or 12-16yo adolescents). Clinical psychology and neuropsychology often aim to draw inferences about people with specific psychological challenges, sometimes aiming to both understand these challenges and what they might apply to the broader population (healthy controls). Some areas are more complex, such as social psychology which looks for behavior that may be general to all humans, but may also be strongly culturally or socially influenced (e.g., stereotype bias). Biologically-oriented psychology disciplines such as health and neuroscience look for generalities but acknowledge that these can be influenced by biological differences such as genetics.

In some of these domains, the research hypothesis provides a very specific approach to identifying participants who can be invited to participate in the research study. However, for a wide range of research topics, participants are recruited broadly from the available population and in these cases, some attention needs to be paid to the sampling procedure to identify if that affects the conclusions of the study. Thus far, most of our research design has been aimed at good operational definitions and control of extraneous variables to avoid confounds to maximize our **internal validity**. The separate question of whether the conclusions from an internally valid study apply broadly to the whole human population is a question of **external validity**.

External validity describes the degree to which the conclusions of the research study can be applied to the rest of the population outside the specific sample who participated. Good external validity means the results apply to the whole population. External validity can be limited if some aspect of the procedure to recruit participants into the sample accidentally introduced some bias such that the sample is no longer representative of the population. Examples of this are studies run exclusively on undergraduate populations, which typically have a very restricted age range (among other characteristics). Inferences drawn about the operation of memory or perception from this range may not apply to all other ages. Some of these limitations are obvious and implicit. Nobody expects studies on memory for reading prose passages to apply to pre-verbal infants. Others are subtler, such as the finding that some visual illusions are not perceived universally across cultures but may reflect the experience of the participants with stimuli related to the illusions.

In some cases, the recruiting (sampling) method used to carry out the research can introduce bias into the results. For example, research on high-risk behavior has to carefully consider how to find participants, e.g., an advertisement for a study on sexual attitudes or behavior may not recruit a sample of participants that is representative of the population. This can be a difficult issue to resolve since we ethically require participants to voluntarily participate in a research study. People who are reluctant to talk about this topic even when their privacy is guaranteed may be persistently underrepresented in those studies and it can be difficult even to assess the size of the problem (which in this case, likely also varies substantially across different social and cultural groups).

The question of sampling methodology is most often considered in the context of non-experimental research, especially survey research and the related area of polling. Here we will introduce the main underlying ideas and methodologies as they can be applied to sampling and generalizability in experimental research. In Chapter 17 we will return to this topic in the context of non-experimental research. Many of the sophisticated sampling methods described here, such as **stratified random sampling** are critical for survey/polling methodology but less commonly used explicitly in experimental research.

## Learning Objectives

1. Understand the differences in different kinds of sampling
2. Define sampling bias in general and non-response bias in particular.
3. Understand how to explain the limitations of the recruiting process using in a research study and how this might affect the external validity of the conclusions.

# Sampling and Measurement

In addition to identifying which variables to manipulate and measure, and operationally defining those variables, researchers need to identify the population of interest. Researchers in psychology are usually interested in drawing conclusions about some very large group of people. This is called the population. It could be all American teenagers, children with autism, professional athletes, or even just human beings—depending on the interests and goals of the researcher. But they usually study only a small subset or sample of the population. For example, a researcher might measure the talkativeness of a few hundred university students with the intention of drawing conclusions about the talkativeness of men and women in general. It is important, therefore, for researchers to use a representative sample—one that is similar to the population in important respects.

One method of obtaining a sample is simple random sampling, in which every member of the population has an equal chance of being selected for the sample. For example, a pollster could start with a list of all the registered voters in a city (the population), randomly select 100 of them from the list (the sample) and ask those 100 whom they intend to vote for. Unfortunately, random sampling is difficult or impossible in most psychological research because the populations are less clearly defined than the registered voters in a city. How could a researcher give all American teenagers or all children with autism an equal chance of being selected for a sample? The most common alternative to random sampling is convenience sampling, in which the sample consists of individuals who happen to be nearby and willing to participate (such as introductory psychology students). Of course, the obvious problem with convenience sampling is that the sample might not be representative of the population and therefore it may be less appropriate to generalize the results from the sample to that population.

Essentially all psychological research involves sampling—selecting a sample to study from the population of interest. Sampling falls into two broad categories. The first category, Probability sampling, occurs when the researcher can specify the probability that each member of the population will be selected for the sample. The second is Non-probability sampling, which occurs when the researcher cannot specify these probabilities. Most psychological research involves non-probability sampling. For example, Convenience sampling—studying individuals who happen to be nearby and willing to participate—is a very common form of non-probability sampling used in psychological research. Other forms of non-probability sampling include snowball sampling (in which existing research participants help recruit additional participants for the study), quota sampling (in which subgroups in the sample are recruited to be proportional to those subgroups in the population), and self-selection sampling (in which individuals choose to take part in the research on their own accord, without being approached by the researcher directly).

Compared with non-probability sampling, probability sampling requires a very clear specification of the population, which of course depends on the research questions to be answered. The population might be all registered voters in Washington State, all American consumers who have purchased a car in the past year, women in the Seattle over 40 years old who have received a mammogram in the past decade, or all the alumni of a particular university. Once the population has been specified, probability sampling requires a **sampling frame**. This sampling frame is essentially a list of all the members of the population from which to select the respondents. Sampling frames can come from a variety of sources, including telephone directories, lists of registered voters, and hospital or insurance records. In some cases, a map can serve as a sampling frame, allowing for the selection of cities, streets, or households.

There are a variety of different probability sampling methods. Simple random sampling is done in such a way that each individual in the population has an equal probability of being selected for the sample. This type of sampling could involve putting the names of all individuals in the sampling frame into a hat, mixing them up, and then drawing out the number needed for the sample. Given that most sampling frames take the form of computer files, random sampling is more likely to involve computerized sorting or selection of respondents. A common approach in telephone surveys is random-digit dialing, in which a computer randomly generates phone numbers from among the possible phone numbers within a given geographic area.

A common alternative to simple random sampling is stratified random sampling, in which the population is divided into different subgroups or “strata” (usually based on demographic characteristics) and then a random sample is taken from each “stratum.” Proportionate stratified random sampling can be used to select a sample in which the proportion of respondents in each of various subgroups matches the proportion in the population. For example, because about12.6% of the American population is African American, stratified random sampling can be used to ensure that a survey of 1,000 American adults includes about 126 African-American respondents. Disproportionate stratified random sampling can also be used to sample extra respondents from particularly small subgroups—allowing valid conclusions to be drawn about those subgroups. For example, because Asian Americans make up a relatively small percentage of the American population (about 5.6%), a simple random sample of 1,000 American adults might include too few Asian Americans to draw any conclusions about them as distinct from any other subgroup. If representation is important to the research question, however, then disproportionate stratified random sampling could be used to ensure that enough Asian-American respondents are included in the sample to draw valid conclusions about Asian Americans a whole.

Yet another type of probability sampling is cluster sampling, in which larger clusters of individuals are randomly sampled and then individuals within each cluster are randomly sampled. This is the only probability sampling method that does not require a sampling frame. For example, to select a sample of small-town residents in Washington, a researcher might randomly select several small towns and then randomly select several individuals within each town. Cluster sampling is especially useful for surveys that involve face-to-face interviewing because it minimizes the amount of traveling that the interviewers must do. For example, instead of traveling to 200 small towns to interview 200 residents, a research team could travel to 10 small towns and interview 20 residents of each. The National Comorbidity Survey was done using a form of cluster sampling.

## **Sampling Bias**

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

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

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

## Online Data Collection

An increasingly popular methodology for psychological research is based on using web sites that provide access to research participants as a service. At the time of this writing, Amazon’s Mechanical Turk (mTurk) is a very popular option. The mTurk service was not originally designed for human participants for research but has been applied to this purpose. Another newer system, Prolific, was built with access to research participants in mind. In either system, research studies are made available to registered users, termed ‘workers,’ who can search through lists of opportunities to do tasks for monetary compensation.

## History of mTurk

The name Mechanical Turk refers to a 19th century scam where a machine was purported to be a chess-playing automaton (robot). In reality, the machine was simply a device that allowed a human (typically shorter than average) hiding within to manipulate pieces and play chess well. The reference was likely chosen by Amazon to reflect the fact that the original goal of the mTurk service was to hire humans inexpensively to do cognitive tasks that could not be completed accurately by artificial intelligence programs. For example, evaluating the accuracy of key words or search terms being related to online postings or determining if photos provided on a site selling cars were actually usable pictures of automobiles (both tasks that AI has gotten much better at in recent years). Their model was to create a marketplace where “requesters” could post “human intelligence tasks” for pay that could be completed by “workers.”

Psychological research can easily be thought of as a “human intelligence task” and as a result, psychology researchers were very interested in the possibility of accelerating the process of data collection using mTurk. For experimental paradigms that can be run online, such as surveys or research protocols implemented to operated within a web browser, mTurk can act as a recruiting site to identify participants who volunteer to participate in exchange for monetary compensation. Given the relatively large population of potential volunteers who can be reached through this marketplace, there is a potential for carrying our research studies at a much more rapid pace than can generally be done with in-person data collection.

However, there are handful of concerns that have been raised about online data collection. First was the question of diversity of the participant pool and whether it might embed bias in those samples (see below). Research on demographics of mTurk participants have tended to show that the pool is potentially more diverse that the WEIRD (also see below) participant samples collected on university campuses. Of more persistent concern is the general difficulty of confidently establishing identity online. As we saw in our initial discussion of research ethics, research with vulnerable populations, such as children, requires more oversight. Exactly how rigorous procedures are at mTurk for ensuring children are not able to make accounts and participate in research is not well known. The site also does not provide guarantees the research available to participants meets IRB oversight or even that it pays a fair amount in compensation for time spent. It seems likely that there will eventually be a move towards an online marketplace system designed for human participant research that is integrated better with standard, ethical research practices.

For the purpose of sampling, it is important to note that some of the concerns that people have about collecting data online (e.g., that internet-based findings differ from those obtained with other methods) have been found to be less of an issue than expected. In the table below (adapted from Gosling, Vazire, Srivastava, John, 2004) three preconceptions about data collected in web-based studies are addressed.

 Table. Some Preconceptions and Findings Pertaining to Web-based Studies

|  |  |
| --- | --- |
| **Preconception** | **Finding** |
| Internet samples are not demographically diverse | Internet samples are more diverse than traditional samples in many domains, although they are not completely representative of the population |
| Internet samples are maladjusted, socially isolated, or depressed | Internet users do not differ from nonusers on markers of adjustment and depression |
| Internet-based findings differ from those obtained with other methods | Evidence so far suggests that internet-based findings are consistent with findings based on traditional methods (e.g., on self-esteem, personality), but more data are needed. |

## WEIRD samples

Psychological research is increasingly developing a sensitivity to the fact that a great many well-known findings about psychology depend largely or entirely on undergraduate participants who are at universities with substantial research programs. University students are already a very restricted demographic based on age and education. In addition, research universities tend to be ones that are more competitive with respect to admissions and therefore reflect populations that have succeeded in that competition. That may bias samples with respect to both individual difference variables and also factors like socio-economic status, which have substantial impacts on student preparation, success, and ability to attend competitive universities.

The acronym WEIRD (Western Educated Industrialized Rich Democratic) has been used to describe the potential sampling issue involved in depending on undergraduate participants. Note that the Western and democratic elements of the acronym reflect the fact that to date, the overwhelming bulk of published psychological research that has been done with populations drawn from the United States, Canada and Western Europe. Acknowledgment of this issue has mainly been used to be more explicit about the demographic characteristics of the participants in research in publication. Some effort has also been made to increase outreach to broader and more diverse communities. This is not a simple problem to solve as research at universities with convenience samples of undergraduates is far easier and therefore less costly to carry out. Making research more difficult or expensive will lead to less science being accomplished, which is not necessarily the goal of broadening our sampling procedures. Online/internet based recruiting holds some promise for improving this, but will still tend to over-represent aspects of the WEIRD demographics as internet access and devices (smart phones, laptops, computers) for carrying out research are more available to relatively wealthier groups.

## Limitations in External Validity

Identifying that a research study is based on a WEIRD sample does not necessarily imply that the results are importantly limited by this fact. To identify a limitation, we need to be able to communicate an alternate hypothesis. Specifically, we should be able to identify a different sample that might plausibly not be expected to show the same behavior as the participants in a research study being reported. For example, in our Experiment 1 study examining the effect of encoding “depth” on word memory, the data were collected from undergraduates in a research methods class. We might note this and worry that it limits our conclusions. However, we would need to identify what different samples might not show the depth effect. There is no existing work that suggests that non-college-attending participants, or older/younger participants do not show the effect of memory enhancement following a study process that connects items to be remembered to existing semantic knowledge. Obviously, participant who cannot read (young children, illiterate) would probably not show the depth encoding effect for word list stimuli. We might also see this as a limitation of the stimuli that could potentially be addressed in future studies.

In many cases, the limitation arising from sampling is completely clear. Research on stereotype bias based on race that are run in the USA are likely to show different patterns of behavior than bias studies carried out in other countries. There may be important commonalities that provide insight into human behavior for all humans on the planet, but the sample context is an important part of understanding the result of a research study run in one location. Psychological research based on attitudes, identity, or moral values are all examples of research that is very likely to be related to the population from which the participants were sampled.

Identifying external validity limitations is typically done by brainstorming as many conditions as you can think of where the effects of the study might not apply, then decide if any of these are important limitations to include in the discussion of your results. Hypotheses about effectiveness of limitations in generalizability usually must be done based on general knowledge of people’s behavior. Our intuitions are often useful here, but expertise within the specific subdiscipline of the research is also very helpful. Obviously, the more experience you have in psychological research, the better your intuition about what sampling issues may be relevant.

External validity judgments can virtually never be made perfectly or with absolute confidence. They may look accurate, but then a new idea about differences across people advances our scientific understanding and modifies previous broad statements. Interval validity of studies, when established, rarely changes when new evidence about the phenomenon at study becomes available. However, the external validity of findings may change as science progresses and new factors and context elements are discovered in subsequent research. Often these advances further refine our understanding of the groups of people to whom the results apply, demonstrating the need to be complete and accurate about the samples participating in each research study.

## References

Lahaut, V. M. H. C. J., Jansen, H. A. M., van de Mheen, D., Garretsen, H. F. L. (2002). Non-response bias in a sample survey on alcohol consumption, 256–260.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2004). Hoboken, NJ: Wiley.

Gosling, S. D., Vazire, S., Srivastava, S., John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires (2), 93-104.

Natala@aws. (2011, January 26). Re: MTurk CENSUS: About how many workers were on Mechanical Turk in 2010? Message posted to Amazon Web Services Discussion Forums. Retrieved from <https://forums.aws.amazon.com/thread.jspa?threadID=58891>

Buhrmester, M., Kwang, T., Gosling, S.D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high quality, data? (1), 3-5.