External Validity

External validity is the confidence you can have in **generalizing** your results or findings across people, situations, and times not included in your study. Before you can evaluate a study's external validity, you must first determine whether the study is intending to generalize its *results* (actual numeric estimates of a population, such as voter opinions) or its *findings* (the conclusions it reaches about the relations between variables, such as the relation between heat and aggression).[1]

Generalizing Results

Generalizing *results* requires that the **sample** in a study (the people selected to be in a study) are very representative of the **population** to which the results are to be generalized. Typically, generalizing results requires some kind of **probability sampling**, defined as a process of obtaining participants in which each member of the target population has a known probability of inclusion in the sample. The major types of probability sampling are simple random sampling, stratified random sampling, and cluster sampling.

Simple random sampling. In **simple random sampling**, each member of the target population has an *equal chance* of inclusion in the study. This type of sampling produces the most representative sample but has very stringent requirements. First, you must have an exhaustive list (meaning that nobody is excluded) of all the population members. This can be very difficult to generate. Phone books often exclude people with cellphone-only phone access, people who are poor or homeless, and people who do not wish their numbers to be listed (as much as 50% of the total in metropolitan areas). In addition, women answer the phone more than men, so a simple phone survey may produce more responses from women. Second, you must guarantee that you will have equal access to all the members of the population. This can be difficult if some members are very busy (e.g., medical doctors) or otherwise unwilling to participate. A standard way of evaluating how well a simple random sample actually represents the population is to calculate the **response rate**, the percentage of people that you contacted who provided valid responses. The lower your response rate, the greater the probability that your sample is systematically biased in favor of those who have the motivation and ability to participate. Simple random sampling can be amazingly efficient at accurately representing population values. You may have noticed sample sizes close to 1,100 for polling results reported in the news. If your study is estimating a percentage, such as the percentage of likely voters who intend to vote for a particular political candidate, and your population is over 100,000 people, you only need a random sample of **1,100** people to obtain an estimate that is within plus or minus 3% of the true population value.

Stratified random sampling. **Stratified random sampling** is designed to produce a sample that is exactly representative of the population along one or more dimensions. For example, a sample stratified by ethnicity would contain proportions of ethnic groups designed to exactly match the population. If a population was 15% African American, then the stratified random sample will be 15% African American. To construct a stratified random sample, the first step is to obtain estimates of the percentage of each group you wish to represent in the general population, for example from the latest U.S. Census. Step 2 is to define the size of your sample, say 1,000 people. Step 3 is to compute how many people you will need to select from each group to produce a 1,000-person sample: simply multiply the percentages you obtained in step 1 by 1,000. Step 4 is to *randomly sample within each group* until you have the number of participants you determined in step 3. Stratified random sampling is useful when particular groups in a population make up a very small percentage. Simple random sampling may omit members of these groups just by chance, but stratified random sampling insures that they are included.

Cluster sampling. Cluster sampling is useful when an exhaustive list of individual members is not available, but a list of groups containing almost all individual members is available. For example, a researcher desiring a representative sample of 9-year-olds may be unable to find a list of all 9-year-olds, but could obtain a list of all elementary schools in an area. **Cluster sampling** involves identifying clusters that contain all population members, randomly sampling those clusters (e.g., randomly selecting 20 elementary schools), and including all the members in those clusters. Cluster sampling could be taken a step further by sampling within clusters (e.g., randomly sampling 20 elementary schools, then randomly

1 In this reading, I make a distinction between "results" and "findings" to make a point about judging external validity, but outside of this reading, those labels are usually used interchangeably.

sampling only 2 classrooms within each school). Cluster sampling is more representative when 1) there are a large number of clusters and 2) the size of each cluster is small relative to the population size.

Generalizing Findings

Most research in psychology is not attempting to generate estimates of population values but is instead attempting to measure the relation between variables. For example, Harry Harlow was a researcher interested in infant attachment. The dominant Freudian view of attachment was that it was based on the infant's dependence on the mother for food, but Harlow believed that attachment was due to a need for comfort. He designed a study in which infant monkeys were separated from their mothers shortly after birth and raised in individual cages with two dolls that resembled adult monkeys. The body of one doll was made of a roll of wire (the "wire mother"), but the body of the other was coated with heavy terrycloth (the "cloth mother"). For some monkeys, a bottle of milk was attached to the wire mother, and for others it was attached to the cloth mother. Harlow compared these two groups in several ways to determine whether food was the primary determinant of attachment or whether some other factor (e.g., softness) was important. He measured how much time the monkeys spent curled up to each mother. He introduced a "scary" stimulus (a stuffed bear that clapped cymbals together) and observed to which mother the monkeys ran. He observed whether the monkeys explored an open area more with the cloth mother or wire mother present. The results were clear: the location of the food had almost no effect on measures of attachment. In each case, the monkeys showed an overwhelming preference for the cloth mother. Douglas Mook (1983) argues that for studies such as Harlow's attachment study, external validity is a moot point because the study authors do not intend to have their *results* generalized beyond the sample. These kinds of studies still have value because what can be taken beyond the lab is *an understanding of the processes studied*. Harlow's attachment study is an example of a study that doesn't have much external validity in terms of its *results* (rates of monkey behavior), but that still contributes to our understanding of attachment because of its *findings* (preference for comfort over food). The monkeys were not representative of all monkeys, and the situation was certainly artificial (not many wire mothers in the jungle). Harlow's findings made an important contribution to our understanding of attachment because they contradicted the prevailing Freudian hunger-reduction theory and instead supported a theory of attachment that emphasized contact comfort. Before you disregard a study's results as useless because the study did not use probability sampling, consider whether the study's *findings* make an important contribution.

External Validity Across People, Situations, and Times

External validity is often discussed in three contexts: people, situations, and times. A study may have high external validity with regard to people (e.g., a random sample of 1,100 likely voters) but poor external validity with regard to time (e.g., the sample was collected in 1960). These three contexts are discussed in more detail below.

People. As mentioned above, generalizing results across people typically requires a probability sample unless the processes under investigation are assumed to be fairly universal. One common concern with regard to samples in psychological research is that they are largely convenience samples of college students. David Sears (1986) reviewed the literature on age-related changes in psychological processes and identified a number of cases where college students may be very different from the rest of the population, especially because of their age. For example, compared to middle-aged adults, college students have a self-concept that is still being defined. As a result, college students often express greater uncertainty in their attitudes and are more persuadable compared to older adults. Estimates of the effectiveness of various persuasive methods that are based on college student participants may thus be overestimates for older adults. Another difference that Sears (1986) identified is that compared to the non-college-educated population, college students tend to be more cognitively oriented and less impulsive. These traits could lead researchers to overestimate the role of cognition and underestimate the role of emotion in their theories.

Another major concern with regard to generalizing across people is the effect of culture. Richard Nisbett (2003) found that people from East Asian cultures (e.g., China, Japan) think and perceive in ways that systematically differ from the ways that people in Western cultures think and perceive. For example, compared to Westerners, people from East Asian cultures pay more attention to background events, are more likely to use situational factors (as opposed to personality factors) to explain people's behavior, and

are more tolerant of (and in fact seem to prefer a degree of) contradiction. Thus, researchers must be wary of generalizing results or findings across cultures without comparing samples across cultures.

Situations. To what degree can the results of a study that is conducted in one setting be generalized to other settings? In general, it depends on the degree to which the setting is representative of other settings. This representativeness can be expressed in two ways: **experimental realism** and **mundane realism**. Experimental realism is the degree to which participants' *psychological experience* of a situation is representative of the experience they would have in other situations. Experimental realism asks: Are participants *feeling* time pressure, or social rejection, or conformity pressure? **Mundane realism** is the degree to which the physical setting in a study superficially resembles other physical settings. If you study gambling in a laboratory setting where no real money is involved, is it safe to generalize your findings to casinos? Laboratory experiments are often criticized as having poor mundane realism, but Kruglanski (1975) argues that experimental realism is more important because what ultimately influences behavior is a person's psychological experience of their environment, not the environment itself. Stanley Milgram's obedience experiment has been criticized as artificial and unrepresentative because it took place in a laboratory. Ironically, this fact makes Milgram's results all the more compelling because we would expect participants to take the experience *less* seriously because they were in a laboratory. The films of participants in Milgram's study clearly show their high levels of anxiety and quickly dispel concerns that the participants did not believe they were really injuring another person. If anything, the rates of obedience in Milgram's studies are underestimates of the obedience you would observe in a more realistic setting where the costs of disobedience are much higher, such as in the military. Ultimately, the degree to which the results from laboratory studies generalize outside the laboratory is an **empirical question** (one that can be settled by evidence). Dipboye and Flanagan (1979) compared the results of 200 laboratory and 300 field studies conducted on the same topics (in industrial-organizational psychology) and found them to be equivalent. Thus, it would be premature to dismiss the findings of laboratory experiments simply because they do not superficially resemble a specific real-world setting.

Times. Do the results of research conducted decades ago still apply? Sometimes. But the only way to know for sure is to attempt to replicate findings when you suspect that changes in the culture may impact a particular phenomenon. Replications of obedience and conformity research generally indicate that rates of these phenomena are relatively stable over time, but replications of sexism and racism research indicate that public statements of race- or sex-based superiority have become much more rare over the last several decades.

Nonprobability sampling. When a sample is obtained without the intention of estimating population values, probability sampling is unnecessary. There are two nonprobability sampling methods you should be familiar with. The first is the most common sampling method used: **convenience sampling**. Researchers using convenience sampling simply sample anyone who is willing and able to participate. This method is easy and inexpensive. It is not appropriate if results are intended to be generalized to the population. For research involving basic physiological processes that are very similar across people (e.g., visual adaptation to low-light conditions), a convenience sample is probably not a problem. A sampling method that offers slightly more representative results than convenience sampling is **quota sampling**. Quota sampling is like cluster sampling, described above, except that there is no randomness in the selection of participants. A researcher using quota sampling defines the percentage of participants meeting certain criteria that are desired (e.g., sample will be *x*% White, *y*% African American, etc.) and then finds a convenience sample of each of those groups. The sample will superficially resemble the population along the dimensions that were selected, but because participants were not randomly sampled, a quota will not provide reliable estimates of the population.

Conclusion

External validity, like internal validity and construct validity, is a matter of degree: all studies have it to some degree. No study is "externally invalid," only less generalizable to particular people, settings, or times. In addition, snap judgments about a study's external validity based on the strategy it employs (lab experiment vs. survey) are unwise. Instead, consider the claims the researcher is making with regard to generalizing *results* (numeric estimates of population values) versus *findings* (general conclusions) and consider how participants were sampled.