WHAT GROUPS DO YOU WANT TO STUDY?

3

Before you can begin, you must ask: Whom should I study? The answer comes directly from your research questions. Suppose, for example, you want to examine the effectiveness of a new approach to teaching expository writing. To supplement individual writing outside of class, each student will write collaboratively with a classmate for one hour during each two-hour class. You hypothesize that collaboration will help students develop skills necessary for revising and improving not only their joint writing, but their individual writing as well.

At first glance, the goal here is to examine *students*. The purpose of the innovation is to improve students' writing skills, and the most direct way to assess improvement is to study the changes in their skills over time.

But identifying a target group with a label as broad as "students" is only the first step in selecting people for study. In which students are you *particularly* interested? All students enrolled in freshman writing classes? Only those who have poor writing skills, because they have the most to gain if the new approach is beneficial? Only those who have good writing skills, because they have the crucial foundation for advanced editing skills? The word "students" is a good characterization of the group you want to study, but it is only a starting point.

Taking a broader perspective, are you interested only in how students respond to the innovation, or also in how instructors respond? Does student collaboration free up some of the instructors' classroom time, enabling them to spend more time with individual students? Or does the need to orchestrate successful collaboration between students absorb even more instructor time, leaving less for individual students? Innovative programs may affect participants other than just the target group. Students enrolled in the freshman composition course may be your primary interest, but perhaps they should not be your sole interest.

A key step in designing research is to clearly specify, *before collecting any data,* which respondents are the focus of your study. In this chapter, we develop several themes to help you make this specification:

- *Provide a rationale for all your decisions.* You should have a clear justification for all decisions about the people you will study. Feasibility is important, but your respondents must be chosen because you are specifically interested in them.
- *Consider issues of generalizability.* Will others see your research as useful to them, or will your results be too specific to a particular set of places, persons, and times?
- *Consider different types of respondents.* Many interesting questions in higher education involve not just one type of respondent, such as students, faculty members, or departments, but the relationships between types of

respondents. How do student-faculty interactions affect students and faculty members? What characteristics of teacher behavior enhance student learning? Expanding the types of respondents will give you another window on the phenomenon under study. But beware: studies involving more than one type of respondent must be designed with special care.

Specifying the Target Population

The first step in identifying whom to study is to specify a *target population*. By doing this precisely, you can select a sample of respondents that is representative of that population. With an imprecise specification, you will never know how useful your results are.

How do you identify the target population? One way is to select a target population because of its *generalizability*. Researchers achieve generalizability by using target populations that include a *wide range of persons, places, and times*. The broader the definition of the target population, the more broadly applicable your results, and the more likely other researchers will see the relevance of your results to their interests. Donald Campbell and Julian Stanley (1963) refer to this feature as external validity—how well the findings of a study apply to external groups.

But broad applicability of results is not the sole reason for selecting a target population. Substantive questions also are important. You must decide on the particular group you wish to study. Whom is the new advising system designed to help? Who is at risk for dropping out? Who could benefit from teacher training? Research projects usually evolve from your observation, intuition, or need-to-know about a specific group. That group is your target population.

Generalizability and substance should be foremost in your mind when specifying your population. But as a practical matter, the target population must be delimited precisely by specific characteristics. You must identify which *persons,* which *places,* and which *times.* We have found that four sets of criteria help to identify a population: (1) inclusion criteria; (2) exclusion criteria; (3) expected effect size; and (4) feasibility.

Inclusion Criteria

The major question you should ask yourself when developing inclusion criteria is: Why? Why do I want to study these particular students? Why do I want to study what happened during this period of time? Why do I want to study what happened in this particular department? *You should have a sound rationale for identifying those individuals eligible to be included in the target population.* If you do not have such a rationale, redefine the target population until you do.

What constitutes a sound rationale? In the abstract, the adequacy of a rationale is in the eye of the beholder. But once your research is completed, the adequacy of your rationale will be judged by your audience, be it administrators, faculty members, policymakers, or even students. A sound rationale is thus one that is logical to your ultimate audience. It does not have to be elaborate.

To illustrate, suppose you are interested in investigating the relationship between financial aid awards to freshmen and the likelihood that a student will complete her degree. Your hypothesis is simple—a student who receives financial aid is more likely to persevere and complete a degree than one who doesn't. And as financial aid increases, so does the likelihood that a student will graduate.

What rationale can be used to specify a time horizon—the beginning and end points for your study? For a starting year, you might choose 1975 because of changes in financial aid policies in the mid-1970s. Financial aid data for students enrolled before 1975 may be noncomparable with data for students enrolled after 1975. Or perhaps you should limit your study to students who enrolled after 1982, the year the Reagan administration introduced dramatic changes in the Guaranteed Student Loan Program. Our point is not that certain years are "correct," but rather that you must articulate a defensible reason for selecting the characteristics that circumscribe your target population.

EXAMPLE: ***Specifying the characteristics of your target population: How long does it take to earn a doctorate?***

Every year since 1938, the National Research Council has conducted the Survey of Earned Doctorates by sending a questionnaire to every person who received a doctorate from a U.S. institution. In addition to asking how long it took the student to complete his or her degree, the questionnaire includes items asking about academic topics (e.g., field of study, undergraduate school), financial topics (e.g., type and amount of financial support), and demographic topics (e.g., age, sex, race, citizenship, marital status).

Jamal Abedi and Ellen Benkin (1987) used these data to examine factors associated with the length of time it took graduate students at UCLA to earn doctoral degrees. Although data were available for every year from 1938 until 1985, the authors restricted their analyses to the 4,225 UCLA students (with complete data) who received their degrees between 1976 and 1985.

Why did they choose these beginning and end points? Abedi and Benkin explain: "We chose to limit our population to this 10 year span for two reasons: (1) During that decade there were no major external changes that would cause students to finish more quickly or more slowly, and (2) some of the items in the Survey of Earned Doctorates relating to the variables we wanted to study were changed in 1976 but have not been changed since that time" (p. 7). This justification blends a substantive rationale with a practical rationale.

Exclusion Criteria

When deciding whom to *include,* you are also deciding, explicitly or implicitly, whom to *exclude.* For example, the broad definition of the target population for the financial aid study excludes some potential participants because it restricts attention to incoming freshmen. With this definition, students who transferred to the school as upperclassmen would be excluded from the study.

Exclusion criteria should be stated as explicitly as inclusion criteria. As with inclusion criteria, exclusion criteria must be supported by a rationale. With the financial aid study, for example, you might argue that transfer students have a different time trajectory from that of incoming freshmen. The rationales for inclusion and exclusion help determine the soundness of your design.

Many beginning researchers incorrectly assume that the exclusion of some individuals from the target population limits the relevance of a study. This simply is not correct. Excluding people makes the choice of target population more focused and deliberate. When this happens, your research is more likely to be successful because extraneous factors, which might vary tremendously in a less controlled target population, are being "held constant." In essence, by excluding some individuals from the study, you can obtain better information, although on a narrower population.

EXAMPLE: ***Excluding subgroups can improve your design: Modeling MBA student performance.***

Richard McClure, Charles Wells, and Bruce Bowerman (1986) studied the predictors of academic performance among MBA students at Miami University in Ohio, using as their initial target population students who began the program during a single semester. They then excluded three small subgroups. First, they eliminated students who had withdrawn from the program after completing only one, two, three, or four courses, arguing that such a small number of completed courses could not yield a good estimate of a student's GPA. Second, they eliminated international students, "because of the unknown impact on performance of learning in a language and culture that is not native and because of the difficulty in reconciling their undergraduate grade point averages and GMAT scores with the corresponding scores for American students" (p. 183). Finally, they eliminated part-time students, because they hypothesized that the academic performance of part-time students would be adversely affected by professional responsibilities. These three criteria led to the exclusion of 37 students from the sample, yielding a reduced sample of 89 students.

The exclusions paid off. Three previous studies of the relationship between undergraduate GPA/GMAT scores and graduate GPA had been inconclusive. But by excluding the three subgroups, the authors were able to detect a moderate relationship. The narrower definition of their target population allowed them to find stronger effects, although their results are now generalizable, in turn, to a narrower population.

Expected Effect Size

Some researchers choose a target population because of the size of the effect they expect to find. Choosing a group for which you expect to find a large effect is not uncommon or unreasonable. For example, you might target a new writing program to students who have weak incoming writing skills. There are two good reasons for such a strategy: (1) if you find an effect for *this* group, future research can see if it holds also for *other* groups; (2) if you do not find an effect for this group, chances are you will never find it for any other group.

When you choose a target population because of an expected effect size, the generalizability of your findings to other populations is not a concern; indeed, generalizability is usually sacrificed. Instead, the goal is to find evidence to support or refute your hypothesis for *some* group.

In what types of target populations are you likely to find large effects? Effects tend to be larger within groups in greater need or at higher risk—in other words, those with the most to gain. A study of the relationship between financial aid and college persistence, for example, might focus on a target population in great need of financial aid—such as students with lower family incomes or smaller savings. For these students, adequate financial aid may be a major determinant of whether they graduate. If a broader cross-section of students were studied, one that included some wealthier students, financial aid might appear to have a smaller effect.

The decision to specify a target population in greater need, or at higher risk, may also direct you to certain institutions or periods of time. For the financial aid study, for example, you might collect data at a college where students come from poorer families. Or you might collect data only after the restrictions on the Guaranteed Student Loan Program were tightened. By limiting your focus in these ways, you diminish the generalizability of your findings. But you trade generalizability for an increased probability of detecting an effect. Limited generalizability may be a small price to pay if you can demonstrate an important relationship. A subsequent study can then determine if the detected relationship holds for other, broader, populations.

EXAMPLE: ***Choosing a target population in which the effects are likely to be large: Helping college women "break the ice."***

Charlene Muehlenhard, Laurie Baldwin, Wendy Bourg, and Angela Piper (1988) investigated the efficacy of a computer program designed to help college women start and maintain conversations with college men. Rather than using all college women as their target population, they focused their energies on a special subgroup—shy heterosexual women. To identify this group, they administered the Survey of Heterosexual Interactions for Females (SHI-F) to 663 women enrolled in introductory psychology classes at Texas A&M University, and then selected a sample of 45 women with especially low SHI-F scores, which indicate shyness.

These 45 women were randomly assigned to one of three groups: (1) a group that used a computer training program designed to help women initiate and maintain conversations with men; (2) a group that read a written training manual with the same goal; and (3) a no-intervention control group. The researchers also selected 15 women with average SHI-F scores as a "not-shy" control group. The 60 women filled out the SHI-F two more times—once immediately following treatment, and once four months following treatment.

Both the computer program and the written intervention worked. The women in these groups had much higher SHI-F scores at both posttest and follow-up than they had at pretest. The women in the two control groups had relatively stable SHI-F scores, stably high for the not-shy control group and stably low for the shy control group. By using a shy target population, the researchers were able to demonstrate the effectiveness of their training program. Had they focused their efforts on all women, their program might have had little effect.

Feasibility

A study, however well designed, will never succeed if it cannot be implemented. Research projects formulated in a vacuum, without attention to institutional policies, practices, constraints, and philosophies, will not get off the ground. Practical issues such as access, rapport, and logistics must be considered carefully when specifying a target population. Carrying out a study without the formal cooperation of an institution and the informal cooperation of its staff is virtually impossible. Cooperation is part and parcel of applied research.

Feasibility is a necessary, *but not sufficient,* condition for choosing a target population. Some researchers use feasibility as the primary rationale for specifying a target population, and ignore more important considerations such as generalizability. This practice is reflected in the many research studies in higher education conducted in one institution or a single professor's class. Some of these researchers have chosen the particular target population simply because their own institution and their own classes provide an easy source of data. This is inappropriate if the students enrolled in a specific college or course are not the *real* target population.

Not all single-institution studies are inappropriate. The effectiveness of a new policy at your institution can be best evaluated at your institution. If your research question is specific to one college, then it is entirely appropriate to use that college, and that college alone, to provide the target population. This rationale underlies much of the internally sponsored institutional research conducted on the nation's college campuses. Our point is that if you want to make a *broader* statement about a policy, you should evaluate it in a more general setting. To answer the latter type of research question, choosing a single institution will not be sufficient.

Tradeoffs among the Different Criteria

Several of the criteria for specifying a target population conflict. If you choose a target population because you expect to find a large effect, you may sacrifice generalizability. If you choose a target population because of its generalizability, you may sacrifice feasibility. How can you reconcile these conflicts? Although no single answer is applicable in all research settings, we have several recommendations.

Taking a sound rationale as a given, what do we think about the other criteria? In general, we believe that generalizability takes precedence. The broader the target population, the broader the statements you will be able to make about the effects you have investigated. You will be able to figure out if the relationships you have found for the group as a whole hold up across subgroups, or if they are weaker in some subgroups and stronger in others.

There are some research situations in which maximizing the effect size, or ensuring feasibility, should take precedence. We recommend choosing a target population in this way when conducting small-scale research in which the success of your study will be restricted by other factors, such as sample size or the practical difficulties of doing field research. This is particularly true during the early stages of a research enterprise, when investigators have yet to demonstrate *any* effect, let alone a generalizable one that holds across colleges, persons, places, and times.

Where Should You Conduct the Study?

Within the United States, there are some 2000 four-year colleges and universities and 1000 two-year colleges. Each college has its own hierarchical structure of divisions, departments, courses, residence halls, and so on. So when deciding *whom* to study, you must simultaneously decide *where* to conduct the study. Over and above the practical questions of feasibility and access, you must make some substantive decisions. In the sections that follow, we suggest several ways to decide where to conduct a project.

Everywhere

To achieve results that are generalizable across the broad sweep of American higher education, you could argue that the "best" study would be conducted using students and faculty members from the entire pool of postsecondary institutions in the United States. If you collect data on a random sample of students or faculty members from these 3000 schools, the results are easily generalized to students throughout the country.

Using this definition, the series of studies conducted by the Cooperative Institutional Research Program (CIRP) are among the most generalizable higher education research projects. Initiated in 1966, CIRP collects data on randomly selected college students at approximately 300 schools, using a national probability sample. The specific institutions vary from year to year, but collectively, each year's data are generalizable to the college student population of the United States for that year. In addition, the CIRP data base includes longitudinal information on approximately 200,000 students at these schools.

The target population of the CIRP studies is broad, and the researchers use well-designed and stringently applied principles of probability sampling to select institutions and respondents. With such large amounts of data available, the researchers are able to compare findings across different types of persons (e.g., men versus women), places (e.g., private versus public institutions), and times.

Many Places

Many research questions can and should be studied using more narrowly defined target populations, such as groups of institutions meeting a set of specific criteria. For example, L. David Weller (1986) studied the attitudes of college deans toward grade inflation: Did deans perceive grade inflation as a problem on their campuses, and what factors did they identify as contributing to it? Instead of surveying the deans of all undergraduate institutions, Weller limited his study to two types of schools: liberal arts colleges and colleges of education. He identified all American schools in these two categories, and randomly selected 205 liberal arts colleges and 100 colleges of education for study. Seventy-five percent of the deans of liberal arts colleges and 71 percent of the deans of colleges of education stated that grade inflation was a current concern on their campus. Because Weller used good random samples, his results are generalizable to all liberal arts colleges and colleges of education.

Selected Places

Generalization requires you to specify a broad target population, and then randomly sample from that population. But often you cannot select a representative group of institutions. For example, the intense requirements of data collection within each school may make it impossible to collect data at more than a handful of convenient schools.

An alternative method for achieving generalizability is to identify a small number of locations or "sites"—schools, colleges, or departments—where you will collect data. Within each site, you collect data on many respondents. When analyzing the data, you determine the extent to which the findings are consistent across sites. Consistent results suggest that findings are generalizable to a broader group of sites, while inconsistent results suggest that findings may be specific to the sites you have studied.

The challenging question then becomes: "Which sites should I select?" Although it might seem that the best solution is to select a few sites at random, such a strategy is usually ineffective. A handful of sites rarely gives a good picture of the entire target population, so a better strategy is specifically to select sites that meet certain criteria. In other words, with only a limited number of sites, consider *purposeful selection*, rather than relying on the idiosyncrasies of chance. Two broad strategies are available for purposefully selecting a limited number of sites: choosing sites that seem "average," or intentionally choosing contrasting, extreme sites.

Average sites. The use of "average," "typical," or "modal" sites has a long and rich history. The Lynds (1926), in their famous study of the patterns of relationships in a community and in families, used a single site, Middletown, as the prototype of a small American town. Medical researchers at Harvard and Boston Universities have, since 1965, studied a cohort of 20,000 residents of Framingham, Massachusetts, to understand patterns of health and normal aging (Dawber, 1980).

The problem with selecting average sites is that it is difficult to identify and defend any particular typical site. Is Oberlin typical? Typical of what? Midwestern liberal arts colleges? How about Louisiana State? Typical of universities? How similar are UCLA and the University of Wisconsin? What precisely is meant by "typical"? If your research question focuses on what *most* students experience, then perhaps large public universities are modal. If your research question focuses on what the public believes a certain kind of student experiences, perhaps small private colleges are modal.

The key point is that *there is no such thing as a typical college, typical department, or typical residence hall.* If you decide to study one or two schools, departments, or residence halls, don't make grand claims of generalizability based on artificial typicality. Generalize your results only to the particular schools or departments you have actually studied.

Contrasting, divergent sites. An alternative strategy for choosing a few sites is to select sites that differ dramatically on characteristics you expect to influence your results. If you find similar patterns of findings across widely disparate sites, there is a reasonable chance they generalize beyond the few locations you have studied. If you fail to find similar patterns across disparate sites, your findings probably do not generalize; they are site-specific.

EXAMPLE: *Using disparate sites to achieve generalizability: Do attrition rates differ by race?*

Jack Bynum and William Thompson (1983) studied racial differences in the rates at which college freshmen persisted until graduation, stopped out temporarily, or dropped out permanently. Because the researchers expected that the educational trajectories of students would differ dramatically by institution, they examined the trajectories for 1120 freshman who entered four small American colleges in the fall of 1977.

To broaden the generalizability of their findings, Bynum and Thompson carefully selected the four schools to represent "sharply diverse educational philosophies, constituencies, students and environmental settings" (p. 41). Although they did not give the names of the schools, to preserve anonymity, they described them in broad outline. College A was a state university serving predominantly white, middle-income students; College B was a state school serving predominantly black, lower-income students; College C was a private university attracting middle- and upper-income students from all over the country; College D was a private school closely affiliated with a small Protestant denomination, whose white and black lower- and middle-income students came from the Southwest.

Attrition patterns differed dramatically by college: Colleges A, C, and D—which were disproportionately white—had substantially higher dropout rates for black students; College B—which was predominantly black—had substantially higher dropout rates for white students. A reader could interpret the inconsistent findings for the two racial groups as suggesting that the dropout patterns were school-specific. But the *consistent pattern* of racial differentials according to the majority or minority status of the racial group at the school suggests a generalizability that would have been missing had a single institution been chosen. Had Bynum and Thompson examined only predominantly white institutions, they would have (incorrectly) concluded that black students are always more likely to drop out.

The authors acknowledge the limitations of a four-site study, concluding their article by saying: "While these findings appear conclusive, the authors extend cautious generalizations beyond these four particular schools. We would welcome replication of our methodology and the reexamination of the same variables in freshman classes at other institutions" (p. 48).

Selecting Your Sample

Most researchers use one of two types of sample selection: probability sampling or convenience sampling. In a *probability sample,* every member of the target population has a known, nonzero probability of being included in the sample. Because all the probabilities of selection are nonzero, every member of the target population has some chance of being included. If the probabilities of selection are the same and independent for all members of the target population, the probability sample is called a *simple random sample.* If the probabilities of selection differ across subgroups of the target population, called strata, the probability sample is called a *stratified random sample.*

Probability samples are a paragon of high-quality research. When you study a probability sample of respondents, you can be confident your results will generalize to the target population from which you chose them. Only probability sampling procedures produce samples that truly "represent" the target population. Most statistical techniques assume that the observations being analyzed are a random sample from a target population. So if you are to interpret the results of subsequent statistical analyses correctly, you should use probability sampling methods.

Nevertheless, many researchers resort to studying *convenience samples.* A convenience sample is just what its name implies—a sample of respondents selected simply because they are easy to get. In a convenience sample, each member of the target population does not have a known, nonzero probability of selection. Some members are more likely to be selected, others are less likely to be selected, and still others have *no chance* of being selected. As a result, convenience samples are not representative of the target population, and results from convenience samples cannot be generalized to the target population. In technical terms, we say that convenience samples are *biased.*

What precisely is wrong with convenience samples? An extreme example illustrates the general problem. A professor wants to evaluate student opinion of her performance in a large lecture course. Rather than administer a questionnaire to a probability sample of students taking her course, she decides to ask all students who come to her office hours during a three-week period in the middle of the semester to fill out her questionnaire. Lo and behold, the students give her high marks for accessibility, openness, and willingness to talk to students. How useful are her results? Not useful at all, because her convenience sample is likely to have been severely biased. Students who come to a professor's office hours have already, perhaps implicitly, decided that the professor is accessible, for if she were not, why bother coming? By involving only those students who come to office hours, the professor is "stacking the deck" in her own favor. The biases in convenience samples are not always so obvious. But because they can be severe, we strongly discourage the use of convenience samples.

Sampling Frames

The first step in drawing a probability sample is to construct a list of all members in the target population. The list need not be elaborate, but it must be complete. It should list *all members of the target population, without exclusions or duplications*. After all, if a respondent is not on the list, she has a zero probability of selection, and this violates one of the crucial tenets of probability sampling.

Developing a sampling frame is one area in which higher education researchers have a great advantage over many other social scientists. Unlike researchers who study broadly defined community-based populations, a higher education researcher is typically interested in a narrowly defined target population—the student body, or the faculty, or the alumni. Detailed lists of the members of these target populations are usually available from the institution, often from routine management records. Registration, payroll, and admissions files, for example, provide ready-made lists of people eligible for sample selection.

Although omissions are probably the most common problem when developing a sampling frame, you should also check for duplications. Duplications can arise when a researcher uses two or more lists to develop a master sampling frame. For example, Seymour Sudman (1976) describes a study of students and staff conducted at the University of Illinois, Urbana-Champaign. From two published directories, one of students, one of staff, an initial sample of 1145 names was selected. Ninety-six names appeared twice, once as students, once as staff, so the sampling frame actually included only 1049 unique names. (Most of the duplicates were graduate students.)

Try to eliminate duplicates from the sampling frame. Otherwise, you may contact some people twice, wasting precious resources and producing an unintended decrease in the final sample size. Duplicate entries also distort the probability sampling mechanism. People listed twice have higher probabilities of selection than people listed once. Because the higher probabilities of selection are unknown, this violates the principles of probability sampling. When duplicates do arise, consult a book on sampling for advice on handling the duplication.

Different Sampling Strategies

With your sampling frame in hand, you can select a probability sample. Several excellent books on sampling describe the details of how to assign identification numbers to units in the sampling frame and then select respondents for study. In this section, we concentrate not on the details of drawing the sample, but on the principles for deciding whether to use a simple random sample or a stratified random sample.

Simple random samples versus stratified random samples. Table 3.1 presents the number of doctoral students at the Harvard University Graduate School of Education in December 1986. The students are classified by their department affiliation: Administration, Planning and Social Policy (APSP); Human Development, Reading and Counseling (HDRC); and Teaching, Curriculum and Learning Environments (TCLE). Across all departments, there are 801 students, with approximately equal numbers in the two largest departments (around 300 in each) and about half as many in the smallest department. Suppose you want to select a probability sample of 80 graduate students. The simplest approach is to select a simple random sample. Sample sizes for two such random samples are presented in the third and fourth columns of Table 3.1. Within the limits of sampling variation, both random samples are representative of the target population of doctoral students at the School of Education who were enrolled in academic year 1986–1987. This representativeness is guaranteed by the principles of probability sampling, and any uncontrolled sampling variation can be automatically accounted for in the subsequent statistical analyses.

TABLE 3.1. DOCTORAL STUDENTS, HARVARD GRADUATE SCHOOL OF EDUCATION, DECEMBER 1986. COMPARISON OF RANDOM SAMPLING, PROPORTIONAL ALLOCATION, AND EQUAL ALLOCATION STRATEGIES.

		Simple random sample		Stratified random sample	
Department	Number in population	I	II	Proportional allocation	Equal allocation
APSP	326	30	36	33	27
HDRC	313	26	34	31	27
TCLE	162	24	10	16	27
Total	801	80	80	80	81

However, each of these simple random samples has a minor problem. Random sample I is disproportionately weighted toward students in TCLE. Although this is the smallest department, the sample (at random) included somewhat more TCLE students than their proportional representation in the population. Random sample II has the opposite problem—it is disproportionately weighted toward students in HDRC and APSP. It contains fewer students from TCLE than their proportional representation in the population. Despite these problems, *because both these samples are probability samples, any ultimate statistical analyses will lead to findings that can be generalized back to the target population from which they were drawn.* But we see that, with simple random sampling, there can be some imbalance in the proportion of respondents selected from each of the departments.

For these reasons, we suggest that you use *stratified random samples.* To select a stratified sample, you divide the sampling frame into discrete groups called strata. In Table 3.1, the strata are departments. In other examples, they might be colleges, schools within colleges, types of students, and so on. Each member of the target population must be classified into one, and only one, stratum. Thus, the strata are mutually exclusive and exhaustive categories.

Two types of stratified random sampling strategies are most common. *With proportional allocation, each stratum's sample size is proportional to the relative size of that stratum in the target population.* As shown in the fourth column of Table 3.1, APSP, the largest department, with 326/801 = 40.7 percent of the target population, would get 40.7 percent of the sample, for a sample size of 33. HDRC, with 313/801 = 39.1 percent of the target population, would get 39.1 percent of the sample, for a sample size of 31. The balance of the sample would go to TCLE, with 20.2 percent and a sample size of 16. Proportional allocation improves upon random sampling by ensuring that the sizes of the samples within strata *perfectly* reflect the sizes of the strata within the target population. Under proportional allocation, the TCLE sample would always have 16 students; no more, and no less.

With *equal allocation,* sample sizes within strata are predetermined to be equal, regardless of the sizes of the strata in the target population. A sample size of 81, for example, would include 27 students from each of the three departments. With equal allocation, APSP and HDRC are undersampled, while TCLE is oversampled. Equal allocation is the probability sampling strategy that ensures you will have sufficient people to answer your questions within each stratum.

When should you stratify? Stratified sampling is most helpful when the distribution of respondents in the target population is unequal across strata, as in our simple example. For example, stratifying by student gender when studying undergraduates at a college with an unbalanced sex ratio helps to ensure an adequate representation of both men and women.

The advantages of stratified sampling diminish when the target population has an approximately equal distribution of respondents across strata. For example, although it is easy to stratify a population of undergraduates by class year, it generally has little value, because most colleges have approximately equal numbers of students enrolled in each class year. So even a simple random sample of students would yield approximately equal numbers of students for each class year unless the sample is very small and erratic. However, *you will never be disadvantaged if you do stratify* providing you have the resources to do the job well. In fact, if information on substantively interesting stratifiers is available, we believe that you should *always* stratify. You cannot lose, and you *may* gain.

When deciding between equal and proportional allocation, you should examine the degree of imbalance in the sizes of the strata in the target population. When the strata are about the same size, the two stratified sampling strategies will yield approximately equivalent results, and so proportional allocation is preferable simply because it is easier. When the strata differ in size, equal allocation is more attractive. By using equal allocation, you can ensure that you will have enough data within each small stratum to be able to examine differences among subgroups.

However, the gains associated with equal allocation come at a cost. Because in equal allocation the number of people within each stratum of the sample is not proportional to the number of people within the corresponding stratum in the population, you must use sampling *weights* in all subsequent statistical analyses. Cases from oversampled strata get smaller weights and cases from undersampled strata get larger weights. Without weights, the sample data would disproportionately represent the oversampled strata. For a more detailed discussion of weighting, and how subsequent statistical analyses are affected by it, see the books by Richard Jaeger (1984) or Seymour Sudman (1976).

More Than One Type of Respondent

Answering some research questions will require data on more than just one type of respondent. For example, studying the improvement in writing among students enrolled in collaborative writing classes only tells you how the *students* respond to the innovation. To understand fully the effects of the writing program and to examine how it could be implemented on a larger scale, you need data from the *faculty* too.

Using several types of respondents reveals a broader perspective and allows you to answer questions about the *relationships* between the responses from different types of respondents. In the writing program, for example, you might examine how faculty techniques and student performance are associated. Do students learn more when the instructor actively encourages collaboration by working with students in class? Or do students learn more when the instructor is passive and leaves collaboration to the students? We call such questions *cross-level* because they focus on the relationships between data collected for different "levels" of respondents. Fundamental questions about how higher education actually functions are frequently cross-level. They often take the form: How do features of the institution and the classroom affect student learning?

When research questions involve several types of respondents simultaneously, deciding precisely whom to study becomes complex. It is harder than specifying a single target population and drawing one random sample of respondents. To study multiple types of respondents, you must specify the target population for *each* distinct type, and develop plans for selecting people from each population. Designing studies with several types of respondents therefore involves taking into account the hierarchical organization of respondents.

In this section, we describe two approaches for designing studies with more than one type of respondent: the selection of *unlinked samples* and *linked samples*. We describe each and outline its strengths and weaknesses.

Unlinked Samples

To select unlinked samples, you independently specify each of the several target populations. In an unlinked study of student and faculty views about academic advising, for example, you would specify two separate target populations: one of advisors, one of students. Each target population would have its own set of characteristics for specifying membership. You might try to make the specifications similar—for example, you might limit both target populations to certain schools or departments within your college—but such a correspondence is not necessary. To select the respondents, you draw two probability samples: one from each of the two target populations.

When different types of respondents are studied using unlinked samples, the data sampled from the different populations cannot be routinely linked together on a case-by-case basis. For example, if you ask each student for the name of her advisor, and you separately ask each professor for the names of her advisees, you will likely be missing detailed advisor data for some students and detailed advisee data for some faculty members.

Because the two samples are not coordinated, you must conduct two separate analyses: one for advisees, one for advisors. You might look for similar patterns, but you cannot compare each advisor's responses with her advisees' responses. Therefore, unlinked samples do not allow you to study cross-level questions.

Collecting data on several types of respondents with an unlinked design is tantamount to conducting independent studies of the same topic in different populations. Each study is designed to be optimal for describing the responses of a specific population. But because responses cannot be linked across samples, you cannot fully capitalize on the different sources of information to answer cross-level questions.

EXAMPLE: ***Using unlinked samples: Attitudes toward advising.***

Gary Kramer, Norma Arrington, and Beverly Chynoweth (1985) conducted an unlinked study of the undergraduate academic advising system at Brigham Young University. Three distinct target populations were identified: students, faculty, and administrators. For each target population, the researchers selected a stratified random sample of respondents: students were stratified by academic level and college; faculty were stratified by academic rank and college; and administrators were stratified by college. Each type of respondent received the same questionnaire, thereby allowing the researchers to compare aggregate responses from each type. For example, they report that "twenty-seven survey items [out of 49] produced significant differences among all subpopulations . . . Students consistently rated survey items lower than did faculty or administrators" (p. 27). Because the samples were not linked, however, the researchers could not determine whether student-advisor pairs were likely to share the same views.

Linked Samples

To draw linked samples, you collect data on "related" respondents with the goal of analyzing the relationships between their responses. In a linked study of academic advising, for example, you would collect data on advisee-advisor pairs, so that you could not only profile responses for advisees and advisors separately but also study the assocation between the two sets of responses.

Linked designs yield much more information than unlinked designs. Because respondents are linked, you can address both single-level and cross-level questions. For example, you can describe not only student views and faculty views but also the relationship between the two. However, this additional information comes at a cost. Designing a linked study of multiple respondents is much more difficult than designing an unlinked study. The major difficulty lies in the selection of the base target population, and the way you subsequently identify the linked respondents.

The crucial question to consider when designing linked samples is: *Which type of respondent should be the base target population?* In the advising study, for example, should you first select a target population of advisees and then collect data on their advisors, or should you first select a target population of advisors and then collect data on their advisees? There is no single correct answer to this question. Different approaches are best for addressing different types of research questions. The problem is that although each approach is optimal for some questions, it is suboptimal for addressing others. Thus, choosing the base target population is a crucial decision. It requires you to decide which types of research questions are most important to you.

To understand the consequences of choosing different base target populations, compare two linked samples for the advising study. Suppose your primary interest is in the advisors' viewpoints. Then you should select a target population of advisors, such as all faculty members in the college of liberal arts. You might stratify this target population by department, and select a proportionately allocated random sample of advisors from each department. Your sample would include more people from departments with larger faculties. It would represent the target population of advisors—faculty members in the college of liberal arts—very well.

How should you select a linked target population of advisees? The best approach is to identify all advisees assigned to each advisor in your sample. Then select a sample from this target population by taking either all advisees assigned to the selected advisors or a random sample of advisees assigned to the selected advisors. The key point is that, either way, the advisee sample is explicitly linked with the advisor sample. This design is excellent for describing advisors, and for comparing their responses to their advisees' responses.

If you are most interested in the viewpoints of advisees, use a different approach. Select a target population of advisees, such as all juniors and seniors in the college of liberal arts. You might stratify this target population by department, and select a proportionately allocated random sample of advisees from each department. Your sample would include more students from the more popular departments. It would represent the target population of advisees—students in the college of liberal arts—very well. To obtain a linked sample of faculty, select the advisors of all the sampled students. This design is excellent for describing advisees, and for comparing their responses to their advisors' responses.

Nonresponse Bias

In Chapter 8, we discuss how big a sample you need to take for a particular project. Yet the best-laid plans can lead to disaster if you fail to reach all the people you target. We have seen response rates vary enormously, from a low response of 2 percent in an alumni survey, to a 94 percent response rate in a study of extracurricular activities and part-time work at Harvard (Angelo, 1989).

The biggest threat to your results when many people in your target sample don't respond is nonresponse bias. You face such bias if the people you reach give different answers, on average, from what those you didn't reach would have told you. Since it is hard to know with any confidence what nonrespondents would tell you, you face an unknown level of bias when nonresponse is high (Hoaglin et al., 1982).

Before we worry about bias, what can cause nonresponse to a survey? Here are some possibilities:

- People are not at home (or at work, or in the dorm, or in class) when the interviewer visits or telephones.
- People are at home but choose not to respond.
- People are unable to respond—the respondent may be ill, or not understand your question.

- People are not found. They have moved, for example, or dropped out of a degree program.

How to Fix It

The best strategy for dealing with nonresponse bias is to work to minimize it at every stage in your survey. In Chapter 9, we argue that a pilot study is a wonderful tool for trying out questions, refining your survey instrument, and even field testing your ability to reach respondents in your target sample. We recommend you use a pilot study to make sure your instrument is clear, and that you will reach the students or alumni or faculty you hope to reach. But even good surveys will have some nonresponse. What can you do to deal with nonrespondents? Here are three steps.

Callbacks. These are common. If you want to do a personal interview with many students who live in campus dormitories, you probably won't get them on the first try. A student may be out. She may be at class. She may be busy with another activity. So calling back will involve going back a second, third, fourth time. If this seems like an extraordinary amount of work for little payback, you will be heartened that the statistician Leslie Kish (1965) has pointed out that, while the first call yields the most responses, the second and third calls often have higher rates of response per call.

Sampling nonrespondents. You can take a small random sample of nonrespondents, and work very hard to track down their responses. In an alumni mail survey, for example, this strategy will be especially effective. You can use personal interviews with a small sample of nonrespondents, and generalize your findings to all nonrespondents. This procedure will reduce bias dramatically.

Replacing nonrespondents. Professional survey organizations use nonrespondents from earlier surveys as replacements for people who are nonrespondents in a current survey. This procedure is especially useful if you conduct repeated surveys and maintain files of nonrespondents from past surveys, as alumni and development offices often do. The idea behind this procedure is that nonrespondents to different surveys at least have nonresponse in common. A good fallback strategy for estimating what a nonrespondent would have said in the current survey is to coax a nonrespondent from an earlier survey to respond this time.

EXAMPLE: ***Reducing nonresponse bias in an alumni survey.***

While abstract claims about nonresponse bias are common, few researchers have examined how common such bias is, or estimated its size. Roseann Hogan (1985) has done this, and her findings are fascinating.

Hogan examined annual surveys conducted on the graduate cohorts of thirteen junior colleges in 1980 and 1981. The two surveys were done in dramatically different ways. The 1980 effort surveyed alumni by mail. No follow-ups were conducted. Cover letters were not included at all colleges. Return envelopes and postage were not provided. The response rate was 35 percent.

In contrast, the 1981 alumni survey was far more intense. Three mailings were conducted. The first consisted of a cover letter, questionnaire, and stamped, addressed return envelope. The second mailing was a postcard reminder. The third was a remailing of all the original material complete with stamped return envelope. The response rate nearly doubled, to 67 percent.

Hogan then asked how responses to each survey compare with known data for all alumni. She found that women are far more likely to respond than men (in both surveys); younger students are more likely to respond than older students (in both); whites are more likely to respond than blacks (in both); and the mean GPA of respondents was consistently higher than mean GPA for all graduates. Just as she expected, the 1981 survey generated responses that were much closer to known population values than did the 1980 survey.

Hogan reports one big surprise. Both surveys found just about identical correlations between variables. For example, the correlations between educational goals and salary, time it took to get a job, and employment characteristics are nearly identical for the two surveys. Hogan concludes her report by speculating that while lower response rates may lead to response bias for estimates of sex, race, income, or GPAs, the bias in estimating correlations between pairs of variables, such as GPA and salary, may be low even when response rate is low.