# WHAT PREDICTORS DO YOU WANT TO STUDY?

# 4

Researchers are seldom interested solely in the average value of the outcome in their sample. They are more likely to investigate *systematic variation* in the outcome. They might ask, for instance: Are men more likely than women to succeed in the study of mathematics? Are students with better school grades more likely than less able students to be satisfied with the college experience? Would a mentoring program increase the chance that freshmen who choose a major in science will graduate? Each of these questions expresses interest in the *connection between a particular predictor* (gender, school grades, membership in a mentoring program) *and an outcome* (success in math, satisfaction with college, graduation). Whether the predictor defines a characteristic of the individual (gender, school grade) or

describes something that is done to the individual (membership in a mentoring program) the question is the same: What is the relationship between predictor and outcomes?

Many predictors are likely to be associated with any given outcome. How can you decide which of the dozens of potential predictors you should concentrate on? How can you ensure that your study will be able to demonstrate a causal link between predictor and the outcome if there is one? How can you decide when a relationship you observe between predictor and outcome is not attributable to other effects that you simply failed to examine? How can you refine your definitions of predictors to maximize your chance of finding potential effects?

In this chapter, we show how you can make these decisions. In your study, you should:

- *Acknowledge that different types of predictors require different strategies for detecting their effects.* Many different predictors may be associated with your outcomes. Carefully select those you are most interested in and design your study with them in mind.
- *Rule out rival explanations for observed relationships between predictors and outcomes.* Some apparent effects can actually be attributable to other predictors you did not examine. Learn how to identify alternative explanations, and how to design your study to eliminate them.
- *Maximize variation in the predictors.* The greater the variation in the predictors, the more likely that your study will be able to detect a relationship between predictor and outcome.
- *Consider statistical interactions among predictors.* The relationship between a predictor and an outcome may not be the same for all people; it may differ according to levels of another predictor. To detect such statistical interactions, you must include both predictors in your design.

## Types of Predictors

The term *predictor* is a broad one. It encompasses all the potential variables that you might relate to whatever outcomes you are studying. Some predictors describe membership in treatments or innovations, such as a new course or a new method of teaching. Others describe respondent characteristics, such as number of hours of students' part-time work, or achievement test scores, or attitudes toward college success. You examine these characteristics as they occur naturally in the population. Any research project may involve either or both types of predictor.

### Membership in an Innovation

Begin with an innovation or a new program. You usually evaluate its effect by comparing what happens to respondents in the program with what happens to those who are not in the program. A new curriculum, new advising system, new financial aid package, or new living arrangement is an innovation.

Categorical variables describing membership in an innovation are popular predictors, and their effects on the outcome are often relatively easy to evaluate. This ease stems from the large degree of *researcher control.* If you design your evaluation of an innovation *before* the program begins, you have control over many aspects of the research—who gets the treatment, how much of the treatment they get, how long they receive the treatment, what else happens to them while they are getting the treatment, and so on. Equally important, you also have control over who does *not* receive the treatment. When you have this control, you can apply it by using a totally impartial mechanism: random assignment. The actual process of random assignment can use coin flips or a table of random numbers.

The big point about random assignment is that an unbiased independent arbiter, *not the researcher and not the study participants*, determines who gets assigned to which treatment. In Chapter 6, we discuss methods of assigning people to levels of a predictor describing treatment, and comment further on the tremendous advantages that random assignment confers on your research.

---

**EXAMPLE: *Innovations as predictors: Evaluating the efficacy of self-paced computer instruction.***

Microcomputer-based instructional programs are enjoying increasing popularity on college campuses. One reason for this popularity is the hope that these computer programs can adapt to individual differences in students' abilities, learning styles, and learning strategies, thereby allowing students to learn at their own pace. Ideally, self-paced instruction should promote student learning.

One hundred first-term freshmen at Ohio State University participated in a randomized experiment conducted by John Belland and his colleagues (1985), on the effectiveness of a self-paced software program for teaching about the systolic and diastolic operation of the heart. The researchers hypothesized that while some degree of self-pacing was good, some *external* pacing, whereby students moved forward in the program regardless of how well they were doing, would be even better.

Students were randomly assigned to one of four groups: three with differing levels of self-pacing and external pacing, and one no-treatment comparison group. The predictors in this experiment were the categorical (or dummy) variables describing the membership of each student in one or another of these treatments. The research question—whether knowledge of the operation of the heart differed by treatment group—simply asks whether the categorical membership variable is a good predictor of student learning.

For the three treatment groups, the researcher's control over the research setting was remarkable:

> The study was conducted during two consecutive days. Four two-hour time periods were blocked off during which subjects reported to

> one of two separate microcomputer based laboratories. All subjects in a particular laboratory at a particular time experienced the same instructional program . . . The groups were not aware of any differences in the programs experienced by the other groups, nor were they aware that requests for feedback and overall time for instructional program completion were being monitored. When subjects completed the instructional program, they were sent to another room to take the five achievement tests . . . All three instructional program pacing groups received the same instructions and received their achievement tests in the same order. (p. 193)

Because the treatment was a one-time use of a computer program, the investigators were able to control each student's experience fully. Students who were not learning the material using the microcomputer program could not go to the library and supplement their learning with a textbook. Nor could they ask their roommates or a professor for help. Students in all groups had identical experiences except for the degree of pacing of their software program. And because of the random assignments, the investigators could be sure that, on average, groups were similar in their abilities, learning styles, and learning strategies.

---

### Respondent Attributes

Characteristics of people, such as their sex, race, date of birth, or year of entry into school, are beyond your control. You cannot change them. Yet such characteristics may be the most powerful predictors of many higher education outcomes, so the value of asking questions about them is high.

Without control over *who* has *what* features, whether a predictor is "truly" related to an outcome is difficult to establish. Critics can argue that other predictors you failed to examine are really what produced the effect. A classic example of such a "spurious" relationship was described by the English statistician George Yule (1926). He found a strong positive correlation between membership in the

Church of England (the predictor) and the annual suicide rate (the outcome). Of course, no one believed this association was causal (Yule himself called it nonsense); it was attributed instead to a third variable—the passage of time—that had simultaneously produced changes in both church membership and suicide.

Critics can also argue that self-selection of respondents into "groups" created nonequivalencies between the respondents and led to your findings. For instance, consider the relationship between years of education (the predictor) and salary (the outcome) for graduates of an economics department. The observed correlation may be *negative*—the more education a graduate has, the lower his or her salary! However, perhaps another predictor—employment sector—has confounded the relationship between education and salary. Graduates with bachelor's and master's degrees generally choose to work in the corporate sector (which pay higher salaries), while graduates with doctorates choose to work in academics or government (which pay lower salaries). On the surface it appears that education is *negatively* correlated with salary, when in fact, controlling for employment sector, the variables are positively associated.

Rival explanations always loom large as alternatives that might explain away such findings. The challenge for designing good research is to presumptively rule out as many rival explanations as possible.

Alternative explanations of relationships among predictor and outcome are not always easy to rule out, and this makes the problem very serious. For example, Rick Schrager (1986) studied fraternity members at the University of Illinois at Urbana-Champaign and found a correlation of .60 between students' college GPA and their reports of the fraternity's emphasis on academic achievement. Did the fraternity's emphasis on academic achievement *cause* students to perform better? Or did the better students select academically oriented fraternities because they thought

they would fit in better? It's difficult to say, but both explanations probably have an element of truth.

---

**EXAMPLE: *Studying respondent attributes: The effect of same-sex and cross-sex role models on the subsequent academic productivity of scholars.***

Elyse Goldstein (1979) reported on the relationship between the gender of doctoral degree recipients, the gender of their advisors, and their later research productivity. Her study provides a good illustration of the problems inherent in examining the effects of respondent characteristics. Goldstein collected data on 110 students who earned their doctorates between 1965 and 1973 at New York University, City University of New York, or the New School for Social Research. Twenty-six were men who had female advisors, 29 were men with male advisors, 30 were women with male advisors, and 25 were women with female advisors. Students with advisors of the same sex wrote more articles (an average of approximately 2 articles during a four-year postgraduate period) than did students with advisors of the opposite sex (who had an average of approximately .5 articles during this period). The effect was consistent for both men and women.

What can we conclude about the relationship between gender of advisor and later research productivity of the students? Correctly, Goldstein interprets her results with caution: "A causal relationship between scholar/advisor sex and academic productivity cannot be inferred. We have no way of knowing whether male and female subjects in same-sex groups would have performed even more successfully had they had opposite-sex advisors" (p. 409). She points out how participant selection bias may have influenced the findings—more ambitious or more able students may have sought out same-sex advisors, hoping to establish a long-term mentoring relationship. As long as students are free to choose their advisors and advisors free to choose their students, selection bias cannot be ruled out. Without random assignment of students to advisors, we will never be able to determine whether findings such as these are attributable to the sex matching of advisors and students, or to other, unexamined characteristics of the advisors and students.

---

## The Important Role of Variation

Bigger things are usually easier to detect than smaller things. This simple principle applies not only to everyday life but to research as well. When a predictor is related to an outcome, the magnitude of the outcome will be detectably different at different levels of the predictor. The stronger the relationship, the larger the differences in outcome among levels of the predictor.

The relationship between part-time work and student performance illustrates this. Before you can conclude that these variables are related, part-time work must have a sufficiently adverse affect on student grades; the grades of students engaged in different levels of part-time work must be detectably different. If the effect of part-time work on grades is dramatic, and is noticeable even among students who only work a few hours per week, a study to detect the effect would be easy to design. For instance, you could compare the GPAs of working and nonworking students, and because the effect is so large, it would be noticeable even with a modestly sized sample. But if the effect of part-time work on grades is small, then a study to detect the effect would be harder to design. You could still compare the GPAs of working and nonworking students, but you might not be able to separate real effects from sampling variation.

The key message is that, unless there is variation in *both* the outcome and the predictors, no effects can be detected. Or conversely, if you ensure large variation in the outcome and the predictors, then any effects that are present are more likely to be detectable. Therefore, you should always plan your study so that, by design, as much variation as possible is built in to both the outcome and the predictors. Or, at least, so that none of the variation that occurs naturally *in the population* is unknowingly sacrificed *in the sample* by poor design.

### Stratifying to Ensure Representative Variation

You can build adequate representative variation into your sample by using stratified sampling (Chapter 3). Whether your predictor is categorical—perhaps describing an innovative treatment—or continuous, the strategy is similar. Simply stratify according to values of the predictor and draw separate random samples within each stratum. Stratifying the predictor ensures that the sample contains a full range of different sorts of students. Drawing the sample randomly from within each stratum then ensures a representative variation in the outcome. Thus, the likelihood that a given effect will be detected is increased.

---

**EXAMPLE:** ***Stratifying to ensure variation: Minority students' involvement on campus.***

Glenda Rooney (1985) used a stratified sample to examine minority students' involvement in minority student organizations at the University of Wisconsin at Madison. Because she expected that student involvement would differ for the specific minority student organizations on campus, she stratified the target population into four groups: Afro-Americans, Asian Americans, Hispanics (Chicanos and Puerto Ricans), and Native Americans. Noting that the sizes of the groups in the target population were very different, and wanting to be sure that she could compare responses across minority groups, she oversampled the smallest stratum (Native Americans) and the combined stratum (Hispanics), to ensure adequate representation of all five minority groups. The stratified sampling paid off. It was the Native American students who had the highest participation rate. Because they were the smallest subgroup, she might have not been able to detect this effect had she not used a disproportionate stratified sample.

---

Sometimes information on the stratifiers is not easily available. In that case, consider using a *screening sample*. First select a large sample from the target population. Then, for each participant, collect information on the stratifiers and stratify the screening sample. Then select a subsample for in-depth data collection.

If the target population is accessible, such as students in your classes, a screening approach may be feasible. But because a screening sample must be much larger than the final sample, screening costs can be high. When a screening sample is not feasible, because of logistics or resources, consider stratifying by an another variable, highly correlated with the one you want. For example, if you want to stratify freshmen by writing ability, instead of collecting writing samples for a large screening sample, use data from English achievement tests. These scores are not perfect predictors of writing ability, but they are sufficiently correlated with it. Stratification based on these tests would ensure representation of people with different writing skills in your sample. The achievement test scores are for sampling purposes only, and measures derived from actual writing samples become predictors in your subsequent analyses.

### Don't Restrict the Range

If, for any reason, the range of values of the predictors or the outcomes are restricted in any way, then their net variability will be decreased and the likelihood of detecting an effect will be diminished. In practice, you can restrict range in at least two ways: (1) by using measuring instruments with coarse, limited scales so that "observed" scores do not adequately represent "true" values; (2) by studying unnaturally restricted or homogeneous samples. The first of these is relevant to all empirical research and can be ameliorated by the building of better and more reliable

measures (see Chapter 7). The second—the problem of enforced homogeneity—is a particular problem in higher education.

Many studies examine the relation between academic qualifications of incoming freshmen and college GPA. In a 1977 report issued by the Educational Testing Service, Susan Ford and Sandy Campos summarize the results of 827 such studies. They found median correlations of .32 (for SAT math), .37 (for SAT verbal), and .52 (for high school record). But when Ford and Campos estimated the same correlation coefficients using a subsample of 84 studies where the median verbal SAT score of the students was 550 or greater, the median correlations dropped to .26 (for SAT math), .33 (for SAT verbal), and .45 (for high school record). Why? Because the more select subsample was more homogeneous and led to a restriction of range in both the predictor and the outcome. A strong relationship in a *heterogeneous* group is diminished in homogeneous subgroups.

Range restriction gets worse as the sample becomes more homogeneous. As a result, single-institution studies, which are limited to students who met the institutional criteria for admission, are especially susceptible. Paul Schaffner (1985) examined the predictive validity of high school academic qualifications among students at Bowdoin College, which since 1970 has not required applicants to submit SAT scores. Approximately 31 percent of students who matriculate at Bowdoin do not submit SAT scores. Using data from the remaining 69 percent, Schaffner examined the relationship between freshman GPA and the same three characteristics examined by Ford and Campos. He found correlations of .28 (for SAT math), .36 (for SAT verbal) and .37 (for high school record). Schaffner comments that although the correlations for the two types of SAT scores were similar to what Ford and Campos found for selective institutions (such as Bowdoin), the correlation for high school record was lower than expected. Schaffner specu-

lates that Bowdoin's optional-SAT policy, which emphasizes high school records, may have inadvertently restricted the range on this variable even more than might be the case at other selective schools.

---

EXAMPLE: ***The effects of range restriction: Does coaching improve students' SAT scores?***

Dean Whitla (1988) describes a study conducted in 1987 among freshmen enrolling at Harvard University. He asked students who took the SAT *twice* to report whether they had attended a coaching program. Students who reported attending coaching programs gained an average of 94 points between tests; students who reported not attending coaching schools gained an average of 67 points. The 27-point difference between the two groups was well within the standard error of measurement, and thus Whitla was unable to decide whether coaching made any difference.

Can we conclude that coaching has *no* effect? This is a complex question, involving issues of statistical power (Chapter 8), generalizability of findings at an institution such as Harvard to other colleges (Chapter 3), and the accuracy of self-reported measures of coaching courses (Chapter 7). But let us ask about restriction of range here. Whitla points out that since students admitted to Harvard are a selected group, it is more difficult to detect an association between their SAT scores, or changes in their SAT scores, and any predictor, such as attending an SAT coaching class. To investigate the true effects of coaching on SAT performance, a research design should include students with a broad range of SAT scores.

---

### Continuous versus Discrete Measures

Many predictors are, by their very nature, discrete and categorical. Either a student receives the new curriculum

or she receives the old one. She majors in the humanities, the physical sciences, or the social sciences. She lives in a sorority house, a dormitory, or off-campus. Each category of these predictors simply names the group to which a student belongs. Variation in the values of these predictors over people is fixed once the levels of the variable have been decided.

Other predictors are measured along an underlying quantitative continuum. Incoming qualifications measures, such as SAT scores, achievement test scores, high school GPA, and high school rank, are all continuous predictors, as are student age, number of credits taken per semester, and amount of financial aid. On no account should perfectly good continuous predictors like these be artificially reduced to a small number of discrete categories. Incoming students should not arbitrarily have their SAT scores classified as "high," "medium," or "low." Amount of financial aid should not be dichotomized into "no aid" and "some aid." The very act of categorization reduces the intrinsic variation in the predictor and throws away information.

In fact, some predictors often treated as discrete should always be reconceptualized as continuous if that is possible. For example, the class enrollment of a single student can be measured categorically (enrolled versus not enrolled). But a better representation might include hours of attendance at class or level of participation, both of which are continuous and consequently much more informative. Participation in intramural athletics is usually treated as categorical (participant versus nonparticipant). But number of hours spent in practice, number of games played per year, and number of sports played are all continuous variables and may better represent the student's commitment to intramural athletics.

Continuous predictors are all around us. But if they are so common, why are so many studies designed to incorporate only discrete predictors? The answer is simplicity. When planning a project, many researchers find it easier

to conceptualize their predictors as discrete so that they can regard their research as comparing the "haves" with "have-nots." You can always improve your research by treating your predictors as continuous variables. The improvement arises not only from being able to design more variation "into" your predictors but also from being able to take advantage of more powerful statistical techniques to handle your analyses (see Chapter 8).

A consideration of the effects of part-time work on academic achievement illustrates this point. Rather than designing a study of part-time work as a comparison between two groups, you should define part-time work as a continuum, measured in number of hours per week. Combining all students who work part-time into a single, monolithic group is inefficient. Because the number of hours worked per week is reasonably easy to measure, it is *the* predictor to use. Maximize the sample variation in this predictor by including in your sample as many students as possible who work very few hours, and many who work lots of hours.

Planning to use continuous outcomes and predictors has a great side benefit: nonlinear relationships between the predictor and the outcome become easier to identify. Suppose, for example, you want to examine the relationship between the number of hours students spend studying and their performance on exams. Literature on preparation and stress suggests that the relationship is nonlinear: for modest amounts of time, the more a student studies, the better her performance; but once a student exceeds an "optimal" number of hours, performance drops. How can you detect nonlinearity? By examining students who have studied for various numbers of hours, you derive a better picture of the relationship between preparation time and exam performance. If you categorize preparation time by grouping students into predefined hourly ranges, then you may uncover a less interpretable pattern.

Just as it makes no sense to artificially categorize a

perfectly good continuous predictor, neither does it make sense to select out groups that are *extreme* on the predictor ("very high" and "very low") for comparison (on the outcome). Unfortunately, this strategy is very common and has an apparently convincing appeal. If you fail to find differences in outcome between these extreme groups, then it is hard to imagine exciting findings emerging from more ordinary comparisons. However, by choosing extreme groups, you are not only throwing away crucial information—the participants who fall between the extremes—with subsequent loss of power, but when you do detect differences in outcome across extreme groups, generalizing your findings can be chancy. Perhaps the students with the extreme values are simply "outliers." You could argue that if they are unusual with respect to the predictors, they might also be unusual with respect to the outcomes. Choosing to study only unusual, atypical cases is logically defective.

---

**EXAMPLE: *Investigating extreme groups: Do low-achieving and high-achieving fraternities attract different kinds of students?***

Fraternities enjoy differing reputations, some emphasizing social life, others community service, and still others academic performance. Roger Winston, Steven Hutson, and Sue McCaffrey (1980) examined whether differences in orientation toward academics accounts for some of the variation in academic achievement across fraternities.

Two extreme groups of fraternities differed significantly with respect to emphasis on academic achievement, independence, and intellectuality. They did not differ significantly on SAT scores. Finding no differences in SAT scores among the two extreme groups, the authors conclude that "differences in academic abilities of members do not account for the differential academic performance of fraternities" (p. 452). But finding differences in social climate measures that one might expect to be associated with

academic performance (academic achievement and intellectuality), they also conclude that in high-achieving houses, "there is not only concern for grades . . . but also concern for intellectual development beyond assigned classroom activities" (p. 453).

How does their selection of the six most extreme fraternities affect our confidence in their findings? Because they found no differences in SAT scores for these extreme groups, it seems reasonable to conclude that, at this university, there is no relationship between fraternity academic standing and incoming SAT scores. This is especially plausible because the relationship between academic standing and incoming SAT scores is unlikely to be nonlinear.

But the use of extreme groups may color your views about differences between high- and low-achieving fraternities. The authors find that 44 percent of the variation in academic achievement scores is attributable to a single categorical variable: fraternity type. Their use of extreme groups may give an exaggerated impression of the relationship between academic achievement scores and academic standing of fraternities.

---

## Other Reasons for Selecting Predictors

Most predictors are included in your study because you want to learn about how they relate to particular outcomes. When you evaluate an innovation, for example, the categorical variable describing membership in the innovative treatment or control groups is the predictor. When you study the relationship between admissions data and college GPA, the admissions data are the predictors. The predictors are the "question variables": they are selected because you think they are substantively interesting or important.

But sometimes you should incorporate other predictors into your design for two other reasons: (1) *to disentangle the effects of the substantive predictors from the effects of other, less-important, "background" characteristics;* (2) *to determine whether two (or more) predictors interact in their relationship with the outcome variable.*

### Predictors as Covariates

Some predictors are measured in a study simply as interesting background characteristics. They are then included in the eventual statistical analyses as covariates, so that the influence of the key substantive predictors on the outcome can be evaluated in the presence of the background effects. When you include covariates in your analyses, you are trying to distinguish the variation in the outcome that is attributable to the covariates from the variation that is explainable by the key substantive predictors. For example, examining the relationship between entering SAT scores and freshman GPA may make more sense when socioeconomic status has been included as a covariate so that the effects of SAT score and background can be separated.

If the covariates and the "substantive" predictors are only weakly related, or unrelated, to one another, this strategy works well. But when the covariates and the primary predictors are strongly related, the teasing out of the separate influences may be difficult, perhaps impossible. To illustrate, suppose you are investigating the relation between athletic participation and scholastic achievement, perhaps represented by college GPA. Does involvement in sports reduce college GPA? The crucial predictor is whether or not a student plays on a team. But many other things affect academic performance, and athletic participation is only one. So to disentangle how athletics affect academic performance, you must include covariates in your analyses.

Now suppose you collect data on just two predictors—incoming SAT scores and participation in athletics—planning to include incoming SAT as a covariate when you evaluate the relationship between athletic participation and college GPA. You will find it difficult to separate out the effects of incoming SAT and athletic participation on college GPA because these two predictors are strongly re-

lated to each other. Athletes often have lower SAT scores than nonathletes, so you often find an inverse relationship between athletic participation and incoming SAT scores. This correlation makes it difficult, perhaps impossible, to evaluate the unique effects of athletic participation on academic accomplishments.

Suppose on a simple graph you plot GPA as a function of both incoming SAT scores and whether or not each student participates on an athletic team. You will probably find two effects. First, students with higher incoming SAT scores have higher college GPAs. Second, athletes have lower GPAs than nonathletes. Can these two effects—the "athlete" effect versus the "SAT" effect—be disentangled? The dilemma is that, because athletes have lower SAT scores than nonathletes, a selection bias exists. Perhaps the poor academic performance of college athletes is attributable not to their participation in athletics but to their weaker academic preparation, which in turn is reflected in lower SATs. Perhaps these students would have low GPAs even if they had not played sports. And perhaps the nonathletes would have just as high GPAs even if they *did* participate in sports. You cannot tell, because you have too little data on athletes who enter well-prepared, and on nonathletes who enter poorly prepared. *The most sophisticated statistical analyses will not enable you to disentangle the effect of incoming SAT from the effect of athletic participation.*

When the effects of two predictors are hopelessly entangled in this way, we say the predictors are *confounded.* In this example, participation in athletics is confounded with incoming SATs. The two groups in our study (athletes and nonathletes) are too different with respect to a background characteristic (incoming SAT scores). Because of this, you are unable to separate the effects, to control for the selection bias.

*Statistical techniques find it very hard to resolve this problem of confounding, for it is a problem of design, not a*

*problem of analysis*. Statistical techniques such as regression analysis will only be successful if the confounding variables are not highly correlated with the crucial predictors you most care about. If a covariate *is* highly correlated with a key predictor, there is no way to disentangle the effects by analysis. No, the problem can be resolved only by design. *You can reduce the high correlation between confounding variables by using the principles of stratified sampling*. For example, you could resolve the confounding of incoming SAT scores and athletic participation by stratifying the incoming class of 3000 students at a large state university by athletic participation (athlete/nonathlete) and combined SAT score (perhaps using three strata: below 900, 900 to 1100, and 1100 or more). This solution is illustrated in Table 4.1.

Notice that membership in one group (athlete versus nonathlete) is highly related to membership in the other

**TABLE 4.1. DESCRIPTION OF PROPORTIONAL-ALLOCATION SAMPLE VERSUS EQUAL-ALLOCATION SAMPLE FROM A TARGET POPULATION: STRATIFYING TO ELIMINATE CONFOUNDING.**

| | SAT score | | | |
|---|---|---|---|---|
| | <900 | 900 to 1100 | >1100 | Total |
| *Target population* | | | | |
| Athletes | 300 | 150 | 50 | 500 |
| Nonathletes | 500 | 1000 | 1000 | 2500 |
| Total | 800 | 1150 | 1050 | 3000 |
| *Proportional allocation* | | | | |
| Athletes | 30 | 15 | 5 | 50 |
| Nonathletes | 50 | 100 | 100 | 250 |
| Total | 80 | 115 | 105 | 300 |
| *Equal allocation* | | | | |
| Athletes | 50 | 50 | 50 | 150 |
| Nonathletes | 50 | 50 | 50 | 150 |
| Total | 100 | 100 | 100 | 300 |

group (low, medium, or high SAT scores). In the target population, the majority of all athletes entering have SAT scores lower than 900, whereas the vast bulk of nonathletes entering is equally divided between the upper two categories of SAT. Your challenge is to select students into your sample to eliminate the confounding.

Suppose you can afford to select a sample of 300 students. You could use proportionate allocation, taking 10 percent of the students from each of the six strata defined in Table 4.1. But unfortunately, proportionate allocation replicates in the sample the confounding of SAT scores and athletic participation in the population. Of the 300 students in a proportional sample, only 5 would be athletes with SAT scores over 1100. Only 15 would be athletes with SAT scores between 900 and 1100. With these small sample sizes, incoming SAT and athletic participation would remain correlated and you could not disentangle the effect of athletic participation from that of incoming SAT on college GPA.

What changes with an equal-allocation sample? Athletic participation and SAT scores are no longer confounded in the sample. Athletes as a group are oversampled, and athletes with high SAT scores are oversampled even more. Nonathletes as a group are undersampled, and nonathletes with high SAT scores are undersampled even more. This over- and under-sampling unlinks the two predictors, enabling you to disentangle the effect of athletic participation from the effect of incoming SAT in subsequent analyses.

### Interactions as Predictors

Researchers mostly ask questions about *main effects*. Does persistence in science differ by sex? By membership in a mentoring program? By the availability of a new financial aid package? However, there is sometimes a different relationship between a predictor and an outcome for different

types of respondents. For example, the relationship between financial aid and persistence in science may not be similar for all sorts of students—it may differ depending on the student's financial need. Financial aid may have a larger impact among needier students than among less needy students. A new collaborative writing program may not be as effective for English majors as for nonmajors. When this happens—when the effect of one predictor, such as financial aid, or an outcome differs across levels of another predictor, such as need—we say that the two predictors *interact.* If you do not design your study to answer questions specifically about interactions, you may not gather enough data in each important subgroup to conduct the necessary statistical analyses later. For example, suppose only one-quarter of the students taking writing classes are not English majors. Then a small random sample of students in writing classes will have few nonmajors. There may not be enough of the nonmajors to allow you to figure out whether the program is particularly effective (or ineffective) for them.

Even when you collect data on many students in each subgroup, values of one predictor can be confounded with values of another. Suppose you want to evaluate the differential effectiveness of a writing program for two subgroups: (1) majors versus nonmajors, and (2) good writers versus poor writers. In the real world, you might expect English majors to have better writing skills than nonmajors. Of course, not all English majors are good writers, nor are all nonmajors poor writers. But suppose the two are strongly related. After collecting data on the writing programs, what will you conclude if the writing of the English majors improved more than that of the nonmajors? That the program is more effective for English majors? Or that it is more effective for students who are initially better writers? Because membership in one category (major) is confounded with membership in the other category (initial writing skill), you will have a difficult time disentangling main effects from potential interactions among these predictors

and the treatment (the writing program) in the prediction of ultimate writing skill.

No amount of statistical analysis can solve these problems. To avoid them, *you should design questions about interactions between predictors directly into your study.* By specifying, in advance, the interactions in which you are interested, you can make sure you have enough students in each combination of predictor values. You can then be sure that when your data are analyzed, you will have enough students in each important stratum to seek out potential interactions.

Putting this recommendation into operation is similar to resolving the problem of confounding variables—you do it by stratified sampling. Stratify by the two predictors you think will interact, and then select separate random samples from each group.

This is precisely what Elyse Goldstein did when selecting her sample of doctoral students to investigate relationships between gender of student, gender of advisor, and later research productivity. Her initial target population was 481 doctoral degree recipients, divided into strata as shown in Table 4.2. By sampling approximately equal numbers of students within each of the four cells specified in the table, Goldstein was able to (1) ensure that she had enough students with female advisors; and (2) ensure that sex of advisor was almost unrelated to sex of student. This equal-

**TABLE 4.2. MATCHING ADVISOR TO ADVISEE BY GENDER: GOLDSTEIN'S STRATIFIED SAMPLING STRATEGY.**

| | | Advisor gender in target population | | Advisor gender in stratified sample | |
|---|---|---|---|---|---|
| | | Male | Female | Male | Female |
| Student gender | Male | 260 | 35 | 29 | 26 |
| | Female | 155 | 31 | 30 | 25 |

allocation sampling prevented the gender of student and advisor from being confounded and enabled Goldstein to examine the interaction between the two predictors.

## The Integrity of Your Treatment

When your key predictors describe participation in an innovative treatment group or a control group, there are additional issues that you must consider when designing your study. You must ensure that each participant receives the same "standardized" treatment. You must plan to monitor and measure the implementation of the treatment, perhaps with a view to accounting for any variation in implementation when analyzing your data and presenting your findings. You must be aware that uncontrollable outside influences may mediate the effects of the treatment, especially when the treatment is of long duration. We discuss each of these issues briefly below.

### Standardizing the Treatment

In an ideal study, everyone in each treatment group has an identical, "standardized" experience. If the treatment is a new advising system, each student should have an advisor who does similar things. For example, each student should have the same number of hours of contact with an advisor. If the "treatment" is part-time work, each student should work a similar number of hours at a job with similar demands. Such levels of control are very difficult for researchers to achieve.

Short-term treatments are probably the easiest to standardize: the briefer the experience, the less opportunity for uncontrolled variation. When you are investigating longer-term programs, treatment standardization is more difficult if not impossible. Suppose you want to evaluate a new

computer-based mathematics course for all entering freshmen who are not mathematics majors. The treatment, of course, is the new program. But this treatment will differ in many subtle ways among students. Not all students who enrolled will attend every class session. Not all students will complete every assignment. Not all students will pay equal attention in class. Not all students will spend equal time on the computer. Indeed, not all students will even complete the course. Although you want to evaluate "the course," different students are actually getting different treatments.

How can such a treatment be standardized? The guiding principle is to work hard to make the stimuli that form the treatment as similar as possible. If different instructors are implementing a new curriculum, make sure they have been similarly trained. Provide each instructor with a reading list, a set of homework assignments, and topics for unwritten assignments. Encourage them to use similar grading criteria. Ensure that each class has approximately the same number of students. If scheduling permits, make sure they meet at roughly the same time of day on the same day of the week. Require every student to spend at least three hours per week logged onto the computer. Higher education is an arena in which such standardization of treatment is often relatively easy to achieve.

The same principle applies to the comparison group: its experiences should also be standardized. The successful evaluation of a treatment's effect depends upon making an appropriate comparison (see Chapter 5). So the experiences of students in a comparison group should also be as homogeneous as possible. When comparing the new mathematics course with the old one, see if any students in the comparison group are taking another mathematics class. Check whether they are taking any related courses, such as computer programming or applied statistics, that may influence their performance. The goal remains the same: standardize your treatments whenever possible.

### Monitoring Implementation of Treatment

Although standardization of treatment is a goal, it is not always attainable. No matter how much you control course materials and training, not all teachers are equally effective. Regardless of pressure to attend class, not all students will attend every meeting. And among those students who do, the degree of engagement will vary. Even in well-controlled studies, each participant will inevitably receive a slightly different treatment.

Because of this, you should *monitor implementation of treatment*. What should you do if you discover differences? Try to eliminate them where possible. Follow all participants—both instructors and students. Are they attending class on a regular basis? Are they fully engaged in the lectures? Are they turning in homework assignments? Are they logging onto the computers every week? You may have little control over their behavior, but you should be aware that the treatment is not necessarily being received by all.

The participants themselves can provide useful information on how well a treatment was implemented. As part of the overall data collection, ask them to reflect on what they felt were the most salient aspects of the treatment. Do they believe they fully received the treatment you were implementing? For example, suppose you want to study the effects on student learning of increased question-asking by instructors. You encourage several instructors in introductory economics classes to ask more specific questions during lectures. The instructors report that they have actually changed their behavior—they are indeed asking more questions during class. But perhaps the students do not perceive it that way. Perhaps they feel that while the professor now asks many questions, she is not open to receiving students' answers. There is a gap between the treatment planned and the treatment implemented. Before plunging into elaborate statistical analyses, you should know precisely what treatment has actually made it to the students.

### Measuring Implementation of Treatment

Another way to resolve the problem of varying levels of implementation is to measure how well a treatment has taken hold and to use this information in your analyses. To this end, even using broad categories is better than assuming everyone received exactly the same treatment. For example, you would be better off classifying students as regular attenders and irregular attenders than assuming they all attended all class meetings. Information on the *amount* of treatment received can then be used during your analysis to figure out whether the amount matters. Was the new mathematics course more effective for students who attended class faithfully? Was it more effective for students who made greater use of the computer?

### Adjusting the Definition of the Treatment

Sometimes implementation of treatment varies so much among participants that you cannot consider all members of the treatment group to be equivalent. If a student enrolls in the new mathematics course but rarely attends class, does she really receive the treatment? If a student signs up for the new advising system but never meets with her new advisor, does she really experience the new advising system?

One way to resolve this dilemma is to use your measurement of the amount of treatment more systematically. By including the *amount* of treatment as an additional predictor in your analyses, you can determine its relationship with the outcome variable.

Another approach is to narrow the definition of the treatment so that only those receiving at least a prespecified amount of it will be regarded as participants. For example, you might restrict the evaluation of the new mathematics program to those students attending at least half the class

meetings. Or you might restrict the evaluation of the new advising program to those students meeting with their new advisor at least once during a semester.

Narrowing the definition of treatment is not especially helpful if it occurs *after* data collection. But if it occurs *before* data collection—during the design phase of a project—it can be valuable for targeting limited study resources. The relationship between part-time work and academic achievement illustrates this. Different students work different amounts of time each week—some work only 5 hours, others as many as 25. It seems likely a 5-hour work week would have a different influence on academic achievement from a 25-hour work week. The question is: Who should be included in the part-time work group?

You could include *all* part-time workers in the treatment group and measure the number of hours each student works per week. You would then use this measurement as a predictor in subsequent data analyses. But suppose your study resources are limited. Then perhaps you should narrow the definition of part-time work to students whose jobs require at least 10 hours of work per week. They would be classified as working part-time only if they met this criterion; the comparison group would still be those students who do not work at all. Within the smaller group of part-time workers, you would still measure the number of hours worked. But narrowing the definition would help you to standardize the treatment, and would ensure that differences between treatment and comparison groups would be clear.

### Outside Influences

Many higher-education programs are long term. They involve semester-long, year-long, or even four-year-long experiences. Because they are long term, each student's experience is not entirely within the investigator's control.

Outside factors can, and usually do, influence the participants. You need to work hard to maintain the integrity of treatment over time, especially if the design includes randomization. When outside factors have not influenced students, you can be reasonably sure that any effects observed in a randomized experiment are a result of the treatment. But when outside factors have influenced students, such conclusions are difficult to draw because the outside influences—not the treatment—may have created the outcomes you have detected.

---

**EXAMPLE: *The influence of outside factors: Modifying medical education.***

The influence of outside factors made it difficult for Rudy Mitchell (1989) to evaluate the Pathways program introduced at the Harvard Medical School in 1985. As part of the Pathways program, students met in small groups with a mentor and used medical cases to learn the traditional medical school curriculum. Among the 82 percent of the incoming class of 1989 who volunteered to participate in the Pathways program, half actually participated and the remaining half were assigned to the older, standard curriculum. But those not in the Pathways program began to hear about what was happening in the program. They formed their own study groups. They borrowed materials used by their friends. What began as a randomized experiment devolved into a messy comparative study, with steadily declining integrity of treatment. By the third year, Mitchell despaired of making valid comparisons, and came to rely more on anecdotes and student self-reports of their experiences.

---

## Choosing Which Predictors to Study

We close this chapter with a question: Among all the predictors available for investigation, how should you select

predictors you will design into your study? Borrowing from the work of Cox (1958), we present an exercise to assist you in identifying predictors worth including in your design.

With your research questions in mind, *make a list of all predictors you think might possibly be associated with your outcomes.* Your list should be extensive, including both obvious and less obvious predictors. It should include predictors you can easily measure and predictors that are more difficult to measure. Conduct brainstorming sessions with colleagues. Try to develop as comprehensive a list of potential predictors as possible. Then, with list in hand, *divide the predictors into four groups.*

1. *Those of great importance.* These are the predictors you are confident will be strongly associated with the outcome or whose effects you are interested in substantively. They are your "question" predictors.
2. *Those of moderate importance.* These are the predictors you suspect will be strongly associated with the outcome, but they are not a primary focus.
3. *Those of moderate importance, but that define small groups.* These are the predictors that you suspect will be strongly associated with the outcome, but which identify very small groups that are not a central interest.
4. *Those of little importance.* These are all the other predictors that might be associated with your outcome.

Predictors in group 1 are your highest priority. Design your study around them. Design your sampling strategy to incorporate these predictors. Predictors in group 2 are important, but not so important that you must design your study around them. These are predictors you will simply measure for all students in your study and possibly incorporate into your analyses. Predictors in group 3 are a special case. Because they define small subgroups, they are typically predictors whose effects on the outcomes you

would like to eliminate. Consider redefining your target population to eliminate these small groups, using the techniques described in Chapter 3. This may diminish the generalizability of your findings, but the complexity of your design and analysis will also be reduced. If the groups you eliminate are small enough, the diminished generalizability will be small relative to your increased ability to pin down the effects of the predictors in groups 1 and 2. Predictors in group 4 have the lowest priority. If you have sufficient resources, collect data on them. But if sufficient resources are not available, simply choose not to measure them. There comes a time in every game when you have to make your bets and place your chips on the table.