HOW MANY PEOPLE SHOULD YOU STUDY?

8

You've stated a research question clearly and reviewed previous research. You've identified the target population and developed a sampling plan. You've thought carefully about predictors and the comparisons inherent in them. You've selected instruments and improved them. You now face a crucial design question: How many people should you study?

When asked this question by colleagues—and we are asked this question more often than any other—we invariably respond "The more, the better." *The more people you include in your study, the better your chances of finding effects that really exist.* But, of course, this advice is too general to be practical. Research is time consuming, and you can't afford to use all your resources collecting data. You need to know not just that "more is better"; you need to know "how many is enough."

In this chapter, we provide guidelines to help you make this decision. We discuss conceptual issues involved in determining the minimum necessary sample size and we give some ballpark estimates of sample size we have found applicable in many research situations. By the end of the chapter, we hope you will:

- *Understand why we say "more is better."* Choosing your sample size is a crucial feature of design. If you don't collect data on enough people, an otherwise well-designed study may not yield statistically significant results, or results clear enough to guide policy decisions.
- *Know how other design features affect decisions about sample size.* The types of instruments you use, their reliability, the types of analyses they support, and expected attrition all affect how many people you should include in your study. Learn how to account for these factors when setting your sample size and how design modifications will allow you to get away with studying fewer people.
- *Get a feel for some ballpark estimates of sample size.* Even without mastering the technical details, you can get an intuitive sense of how big a sample is needed in many research settings.

Why Is Sample Size So Important?

To understand why sample size is so critical, it helps to think through exactly what you are doing when you analyze your data. If you've selected your sample from a clearly specified target population using probability sampling methods, you can be reasonably sure, *within the limits of sampling variation,* that what you find in the sample holds in the population—that your results can be generalized. When you generalize from the sample to the population, you are making a *statistical inference.*

Statistical inference is actually a four-step process leading to proof by contradiction. The first step is straightforward—state your research questions as research hypotheses, statements of the way you think things really are in the population. For example, when studying the effects of using computers to teach Russian, your research hypothesis might be that computer-aided instruction (CAI) is better than traditional "chalk-and-talk" methods. Specifically, you might hypothesize that students taught using CAI methods have higher Russian achievement test scores, on average, than students taught using traditional methods.

Second, reframe your research hypotheses as *null hypotheses,* statements of the way you think things *aren't* in the population, statements you might like to *reject* on the basis of sample data. In the Russian example, your null hypothesis might be that, in the population, students taught using CAI methods have Russian achievement scores *equal,* on average, to those of students taught using traditional methods. You don't really believe the null hypothesis; it is a straw man to be shot down. You hope the data will refute it, thereby supporting your CAI innovation.

Third, using an appropriate statistical test, determine how likely it is that you would have gotten the sample results you did if the null hypothesis were really true. That's what *p-values* tell you—the probability that you would have gotten a result as extreme as (or more extreme than) you actually did, if the null hypothesis were true. In a way, *p*-values tell you how closely the observed data match what you would have expected to find if the null hypothesis were true: if the observed data are inconsistent with the null hypothesis, the *p*-value is near zero; if the observed data are not inconsistent with the null hypothesis, the *p*-value is far from zero (and close to one).

Fourth, use the *p*-value to make an inference, reasoning as follows: If the *p*-value is near zero, the observed data are inconsistent with the null hypothesis, so the null hypothesis must not be true, and you reject it. Rejecting a statement of no effect implies a conclusion that there *is* some effect—that teaching method makes a difference. If the *p*-value is far from zero, the data are not inconsistent with the null hypothesis, so the null hypothesis *may* be true, and you can't reject it. You simply don't know whether teaching method and Russian achievement are or are not related. When a *p*-value is far from zero, it is telling you that, with the sample data you have, you can't answer your research question.

This four-step procedure forms the cornerstone of deductive empirical research. But it does have an inherent drawback: *you can never be* sure *your inferences are correct.* Because you do not study all students in the population, sampling idiosyncrasies can distort your results. You are making an informed guess based on limited evidence from a representative group of students. If sampling variation misleads you, you may be wrong. Sampling variation adds uncertainty to all statistical inference, for and against null hypotheses. Inferences are based on probabilities. *You reject a null hypothesis when you are reasonably sure it is false; you fail to reject it when you can't be sure it is false.* You are never certain; at best, you are very confident.

Kinds of Mistakes

Two types of mistakes are possible. You can reject the null hypothesis when it is really true, making a *Type I* or *alpha error,* or you can fail to reject the null hypothesis when it is really false, making a *Type II* or *beta error.* If CAI is really *not* more effective than traditional methods but *you say it is,* you are making a Type I error. If CAI *is* really more effective than traditional methods but *you don't say it is,* you are making a Type II error. The possibilities of such failures of inference—rejecting a null hypothesis that is really true and failing to reject a null hypothesis that is really false—will always remain with us. The best you can do is try to minimize them.

Type I errors are serious: no one wants to say an effect exists when, in fact, the opposite is true. To minimize the chances of such errors, you test null hypotheses at prespecified *alpha levels,* such as .01 and .05. Conducting tests at low alpha levels doesn't *eliminate* the chance of making a Type I error, it just limits it to a comfortably small value. The most popular alpha level is .05, but this value is not absolute; it is simply a compromise between making a Type I error and never rejecting the null hypothesis at all. Most researchers feel secure in knowing that, with an alpha level of .05, they have only a 5 percent chance of rejecting the null hypothesis incorrectly.

Type II errors are also serious: if an effect exists, you want to have a good chance of finding it. Yet most researchers consider Type II errors less consequential, arguing that, if an effect exists, failing to find it in any one study is not too serious because eventually *someone* will find it! We disagree; the one study you have the biggest investment in is *your* study. Because you want your study to be able to say something definitive, *you must not shrug off Type II errors.* Only when their chances of occurring are low are you likely to find effects that really exist, allowing you to answer your research questions. When the chances of Type II errors are high, you face a dilemma if you are unable to reject your null hypothesis—you will not be able to say whether an effect does or does not exist. You court the risk that, after investing all your time and effort, your research questions will remain unanswered.

How can you minimize the chances that Type II errors will occur? One way is to test your null hypotheses at relaxed alpha levels, say .10 and .15. Using a relaxed alpha level makes you more likely to reject *all* null hypotheses, including ones that *should* be rejected, thereby decreasing the chances of a Type II error. But, of course, you are also more likely to reject null hypotheses that *should not* be rejected, thereby increasing the chances of a Type I error. Although using a relaxed alpha level does decrease the chances of Type II errors, this amounts to little more than statistical sleight of hand.

Type II errors should be minimized in another way—by design. The single most important design feature affecting the occurrence of Type II errors is sample size—*the more students you study, the lower your chances of making such errors*—but other features, such as the precision of your measures and attrition in your sample, also play a role. By making judicious design decisions, you can hold the probability of a Type I error to .05 or .10, while still minimizing the chances of a Type II error.

Statistical Power Analysis

The process of determining how many students (or faculty members) to include in your study in order to control the chances of a Type II error is known as *statistical power analysis*. Statistical power is defined as one minus the probability of a Type II error, and it is *the probability that you will detect an effect that is really there*. By increasing power, you decrease the chances of making a Type II error and increase the chances of finding real effects. If CAI methods are really better than traditional ones, you stand a better chance of finding out.

In theory, simply decide how much power you want and set your sample size (and other design features) accordingly. If you think a 20 percent chance of Type II errors is tolerable, design your study to have a power of .80; to be more sure, design your study to have a power of .90. But increased power comes at increased cost—*you increase power by including more people in your study*. To increase power to .99 usually requires so many people as to be impractical (often several thousand), whereas powers from .70 to .90 can be had with more manageable sample sizes (often from one hundred to several hundred). Although there is no consensus about the power you should routinely adopt when planning your study (as there is with the .05 alpha level), we recommend that you design your study to have *at least moderate power,* between .70 and .90. This limits your chances of making a Type II error to a tolerable level, from .30 to .10, without breaking the bank.

Four factors directly influence the sample size you will need to attain the level of statistical power you have chosen: (1) the minimum effect size you want to have a good chance of finding; (2) the statistical analyses you will use; (3) the precision of your measures; and (4) how many students will drop out after the sample has been selected. In the following four sections, we discuss these factors, show how they are related to decisions about sample size, and provide some ballpark estimates of sample size for different types of studies you might design.

EXAMPLE: *Designing a study with good statistical power: Do college admissions decisions differ by an applicant's race or gender?*

Elaine Walster, T. Anne Cleary, and Margaret Clifford (1970) were among the first researchers to attend specifically to the concept of statistical power while designing a higher-education research project. To investigate whether the gender or race of an applicant affects college admission decisions, they conducted an ingenious experiment. They took the college applications of three real high school seniors in Wisconsin (with three very different levels of academic achievement as measured by high school grades and ACT scores), and systematically manipulated the students' reported gender and race. For each of the three students, *four* different applications were created: one making the student a black male, one making the student a black female, one making the student a white male, and one making the student a white female.

The researchers then randomly selected a sample of 240 colleges from *Lovejoy's College Guide* and sent each of the 12 applications to 20 randomly selected colleges from this sample. By looking at the variation across the 240 admissions decisions, the researchers hoped to ascertain whether males were preferred over comparable females, and whether black applicants were preferred over white applicants—reasoning that, after all, the three sets of 80 applications were identical except for reported gender and race. The researchers did *not* find the effects they expected: although for the "low ability" application males were preferred over females, there were no statistically significant differences in admissions decisions according to the applicant's race.

How much faith can we place in Walster, Cleary, and Clifford's results, especially the finding of no difference by race? We believe the results are especially compelling because the researchers studied so many schools, making it difficult to argue that the null findings might be a consequence of low statistical power. The researchers address this very point directly, noting that the

> sample size can markedly affect the probability of obtaining statistical significance . . . [We specified] magnitudes of effects that are either important or unimportant and control[led] the probabilities of making correct decisions by solving for the sample size . . . In this study we decided that a mean difference relative to underlying variability of 0.5 would be important to detect with a probability of .90. In addition, alpha was set at .05. Specifying these parameters led to the choice of a sample size of 240. (p. 238)

Had Walster, Cleary, and Clifford studied only a few schools, their null findings with respect to race might easily have been attributed to low statistical power. With such a large sample size, however—240 colleges—we find the authors' argument compelling that either there are no differentials by race, or if there are such differentials they are small in magnitude.

What Size Effect Do You Want to Detect?

In Chapters 2 and 4, we introduced the idea of effect size and discussed why bigger effects are easier to detect than smaller effects. If you are searching for large effects, and they really exist, the null hypothesis is *so wrong* that you can see just how wrong it is by studying only a few people. If CAI methods are really so much better than traditional ones, even a small study will reveal the difference. But when you are searching for small effects, even if they really exist, the null hypothesis of no effect is so close to the truth that you must include many students in your study before being able to reject it. After all, if the null hypothesis is *nearly* true, it *should* be difficult to reject, even if you study hundreds of students.

So before determining how many students to include in your study, you must decide how big an effect you want to find. Although this may seem like putting the cart before the horse—if you already know the effect size, why do the study?—it actually is not. When specifying an expected effect size, you are simply indicating *the minimum effect size you consider worthy of your time.* You are deciding on the smallest effect of computer-aided instruction you care about. Is a difference of 5 points in Russian achievement big enough to warrant your interest, or are you interested only in differences of at least 15 points?

It is difficult to decide just how large an effect you care about. But rough guidelines are available, and by using them carefully you can come to a reasonable working decision.

Some helpful advice is given by Jacob Cohen (1988), who provides three rules of thumb:

- A *small* effect is undetectable by the naked eye: a difference of .20 standard deviations between two group means, a correlation of .10 between a predictor and an outcome, or the difference between 50 and 45 percent. A small effect corresponds to the mean difference in heights between 15- and 16-year-old girls—two groups that differ, but not by much.
- A *medium* effect is large enough to be detected by the naked eye: a difference of .50 standard deviations, a correlation of .30, or a difference between 50 and 35 percent. A medium effect corresponds to the mean difference in infant mortality between blacks and whites in the east south central states.

- A *large* effect would not be missed by even a casual observer: a difference of .80 standard deviations, a correlation of .50, or a difference between 50 and 25 percent. A large effect corresponds to the mean difference in height between 13- and 18-year-old girls.

Cohen's guidelines are widely accepted by empirical researchers, and you may find them useful if you have no other information to go on.

A better way to decide on a minimum effect size is to think about *practical significance,* the real-world meaning you can give to effects of various sizes. Practical significance is very different from statistical significance. If you include enough students in your sample, for instance, a difference of 5 points on the SAT will become statistically significant, but for an individual student, an admissions officer, or even a researcher, this difference is probably trivial. A difference of 50 SAT points is another matter.

Practical significance is in the eye of the beholder. You must know your outcomes and how big an effect your predictors are likely to have in relation to them. Because practical significance depends upon the research context, only *you* can judge if an effect is large enough to be important. Don't waste time worrying about minuscule effects; *design your study so that it is powerful enough to detect effects of practical significance.* After all, if an effect is so small that it is barely detectable by the naked eye or an expert judge, should you be spending your time studying it?

A third way to decide on the minimum effect size of interest is to use your research review, especially if you have conducted or have found a meta-analysis. In a meta-analysis, an effect size is estimated for each study; taken together, the distribution of estimated effect sizes gives a rough indication of what the next study is likely to find.

Meta-analyses often reveal a sobering fact: effect sizes are not nearly as large as we all might hope. Table 8.1 presents average estimated effect sizes from six meta-anal-

TABLE 8.1. A SUMMARY OF TYPICAL EFFECT SIZES: MEAN EFFECT SIZES IN SIX META-ANALYSES OF HIGHER EDUCATION.

Topic and author	Number of studies in review	Mean effect size*
Financial aid and persistence (Murdock, 1987)	46	.13
Computer-based teaching (Kulik, Kulik, and Cohen, 1980)	59	.25
Programs for high-risk students (Kulik, Kulik, and Schwalb, 1983)	60	.27
Student feedback on instruction (Cohen, 1981)	22	.38
Coaching for non-SAT aptitude tests (Kulik, Bangert-Drowns, and Kulik, 1984)	24	.43
Keller's personalized system of instruction (Kulik, Kulik, and Cohen, 1979)	75	.49

*Standardized mean difference (see Chapter 2 for details).

yses in higher education, on topics ranging from the effects of programs for disadvantaged students to the relationship between student feedback on instruction and teaching performance. All six meta-analyses concluded that the average effect was different from zero—that the outcomes and predictors were related (that treatment and control groups differed)—but the average effect sizes were in the small to medium range. Because small to medium effects are the norm, make sure your study has enough power to detect them. Only then will you be able to do credible research.

Once you have decided on the smallest effect that interests you, it's easy to figure out how many students you should include in your study. Several books can help you with computational details (see, for instance, Cohen, 1988; or Kraemer and Thiemann, 1987). In this chapter, we simply present some ballpark estimates of sample size that you may find helpful.

Table 8.2 presents the total sample sizes needed to detect "small," "medium," and "large" effects at three levels of statistical power (.70, .80, and .90). Sample sizes are presented for the two major ways of denoting effect size: a correlation coefficient (applicable when examining the relationship between a continuous outcome and a continuous predictor) and a standardized difference between group means (applicable when comparing outcomes between two groups). In all cases, we assume that two-tailed statistical tests are being conducted at the .05 alpha level.

Small effects are difficult to detect. Regardless of the type of effect you are studying and the amount of power you want, you must study several hundred or a thousand students to have a reasonable chance of detecting them. But don't be dismayed. You may never want to design a study to detect small effects because they are not usually of much practical significance.

Medium-sized effects, in contrast, can be detected with a moderate-sized sample, usually between 100 and 200, depending upon the power you want. One popular guideline is that you should include enough people to have a reasonable chance (power of .80 or so) of detecting medium-sized

TABLE 8.2. HOW MANY STUDENTS SHOULD YOU SELECT? SOME BALLPARK ESTIMATES OF TOTAL SAMPLE SIZE.

Type of effect size	Statistical test used	Statistical power	Anticipated effect size		
			Small	Medium	Large
Correlation coefficient	Pearson correlation	.90	1,047	113	37
		.80	783	85	28
		.70	616	67	23
Standardized mean difference	Two-group t-test	.90	1,052	170	68
		.80	786	128	52
		.70	620	100	40

Note: Two-tailed test, alpha = .05.

effects. This allows you to strike a balance between the detection of tiny effects and blockbuster effects, while still keeping your budget in check.

Large effects are easy to detect, even using small samples. If you were comparing achievement scores among students using two different computer-based curricula, for example, you would have a 90 percent chance of detecting differences between the groups with as few as 68 students altogether (34 students per group).

Many of our colleagues examining these ballpark estimates of sample size think about detecting only large effects. They consider designing a study with 20 or 30 students, supporting their decision with lofty talk of practical significance. Don't fall into this trap. Few important effects are actually that large, and if your study has power to detect only large effects, you have little chance of finding the more realistic small and medium-sized ones. After the data are in, and you cannot reject your null hypothesis because your sample is too small, you will have simply missed an opportunity.

EXAMPLE: *Making a preliminary calculation of sample size: The effectiveness of mastery learning systems for teaching calculus.*

Samuel Thompson (1980) conducted an experiment at the U.S. Air Force Academy comparing the calculus achievement scores of students taught using conventional lecture-discussion-recitation (LDR) strategies and individualized mastery (IM) strategies. He stratified 840 freshmen into four ability groups based on their high school GPA and college admissions test results, and within each stratum he randomly assigned equal numbers of students to the two teaching methods.

Thompson's excellent study is well worth reading. He paid careful attention to many methodological details. For example, comparability between the two instructional groups was enhanced by scheduling classes at the same time of day, randomly assigning professors to teaching methods, and having both groups use the same textbook. Observer bias was controlled by having each exam graded blindly by two instructors, one from each instructional group.

But why did Thompson study 840 students? Examining Table 8.2, we see that with that many people he had enough power (between .80 and .90) to detect even small effects—a difference of .20 standard deviations between the two groups. He detected no statistically significant differences in calculus achievement between the two groups: "with the same level of instructional effort, individualized mastery instruction and conventional instruction produced indistinguishable results in mathematics achievement. This result emerged from an experiment in which the design, methodology, and statistical power were sufficient to detect achievement differences of any practical significance" (pp. 371–372). Because he designed a study with a good chance of detecting even small effects—that is, because he studied 840 people—Thompson's findings are especially persuasive.

What Type of Analysis Will You Use?

Statistical power is actually a property of an analytic technique and a corresponding statistical test, *not* of research design. Most hypotheses can be tested in several ways, and some statistical tests are intrinsically more powerful than others. A more powerful test allows you to detect effects of identical size in smaller samples. This means if you can answer your research questions using more powerful tests, you can get away with studying fewer people.

So before deciding on a final sample size, you must think about how you will analyze your data. In general, we have avoided discussing analytic dilemmas in this book, because they often are tangential to the development of good design and considering them would complicate matters. We raise the topic now because, when statistical power is under discussion, analysis *becomes* a design issue.

The sample sizes presented in Table 8.2 assume that you will use simple parametric analyses—Pearson correlation coefficients for examining the relationship between continuous predictors and outcomes, and two-sample t-tests for testing differences between group means. If you think that you will use other analytic techniques, you must modify your target sample size accordingly. We turn now to two fundamental choices that directly affect power and sample size: the use of analyses that incorporate covariate information, and the use of parametric versus nonparametric tests.

Including Covariates in Your Analyses

In Chapter 4, we described the important role of covariates: predictors not of direct substantive interest but likely to be associated with the outcome. In a study of the effectiveness of different ways of teaching calculus, for example, scores on a calculus pretest, or on the mathematics portion of the SAT, might be important covariates. In a study of the impact of athletic participation on college GPA, high school GPA might be an important covariate. Covariates are predictors that you expect to be related to the outcome, and whose impact you would like to disentangle from the impact of the predictors in which you are really interested.

Covariates can be incorporated into your data analyses as extra predictors in multiple regression analysis and analysis of covariance. Including extra predictors in this way enables you to *increase* statistical power. Adding a predictor means using more information; more information means more power. With more powerful analyses, you can study fewer students or faculty members and still detect effects of the same size, or you can study the same number of people with higher power.

Table 8.3 presents the smaller sample sizes needed when

TABLE 8.3. HOW MANY STUDENTS SHOULD YOU SELECT WHEN COVARIATE INFORMATION IS AVAILABLE? BALLPARK ESTIMATES OF SAMPLE SIZE, ADJUSTED FOR COVARIATE INFORMATION.

		Anticipated effect size					
		Small: (correlation =)		Medium: (correlation =)		Large: (correlation =)	
Statistical method	Statistical power	.20	.50	.20	.50	.20	.50
Multiple regression	.90	998	778	103	79	32	24
	.80	742	578	77	59	25	18
	.70	590	460	61	47	20	14
Analysis of covariance	.90	1,010	594	164	128	66	52
	.80	756	444	122	96	50	38
	.70	594	350	96	76	40	30

Note: Assuming a two-tailed test, alpha = .05, and that the covariate and other predictors are uncorrelated.

the correlational analyses and t-tests of Table 8.2 are replaced by multiple regression analysis and analysis of covariance. Sample sizes are given for two situations: when the correlation between the covariate and the outcome is .20, and when it is .50.

Comparing parallel entries in Tables 8.2 and 8.3 shows just how helpful covariate information can be. Even when the relationship between the covariate and the outcome is fairly weak (such as a modest correlation of .20), you can reduce your target sample size by up to 15 percent, depending upon the effect size you are looking for, the amount of power you want, and the type of analysis you anticipate. When the association between the covariate and the predictor is stronger (such as a correlation of .50), you can reduce your target sample size by as much as 40 percent.

The more covariate information you can include, the more power you gain. The sample sizes given in Table 8.3 assume that you are including only one covariate, and that it has a correlation of .20 or .50 with the outcome. Since both multiple regression and analysis of covariance allow you to include as many covariates as you want, you can increase your power (or decrease your target sample size) by including additional covariates. For example, if several covariates jointly predict 50 percent of the variation in your outcome, you can cut your target sample size in half.

But the gains in power (or reductions in sample size) derived from the use of covariate information are realized only if you use good covariates. Choice of covariates is largely a substantive challenge—there should be a compelling reason for including the covariate when disentangling the effects of other predictors. But over and above these substantive issues, good covariates should also meet two statistical criteria: they should be highly correlated with the outcome, and relatively uncorrelated with each other (so that they are not redundant in their prediction of the outcome). By using covariates that meet these criteria, you can gain considerable power.

Parametric versus Nonparametric Tests

Just as you can *increase* statistical power by adding information to your analyses through covariates, so can you *reduce* statistical power by setting aside information. Although this may seem a foolish thing to do—why would you ever want to reduce power?—it is exactly what happens when you use *nonparametric* statistical techniques such as Spearman's rank-order correlation, the Wilcoxon test, the Kruskal-Wallis test, or contingency-table techniques to analyze your data.

Why are nonparametric and contingency-table techniques less powerful than their parametric counterparts? The reason is simple: they ignore important information. Nonparametric techniques replace continuous scores with ranks; contingency-table analyses ignore even the ordering among people, basing results only on the way people are spread out among categories. These substitutions diminish the amount of information contained in the specific data values, leading to reductions in variability and ultimately producing a decrease in power. Parametric techniques such as correlational analysis, multiple regression analysis, and analysis of variance and covariance are intrinisically more powerful simply because they exploit *all* available information in continuous data.

If parametric techniques are so much better, why does anyone ever resort to nonparametric and contingency-table analysis? The reason is that the increased power of parametric techniques comes at a price: parametric analyses require stringent distributional assumptions. In fact, it is the building in of these assumptions that adds information to the analyses. The assumptions differ across analytic techniques, but one common assumption is that, in every possible subgroup of the population of students or faculty members, the outcome must be normally distributed. If assumptions like this are met, parametric analyses are indeed more powerful. But if the assumptions are not met, the differential advantage of parametric analyses disappears, and they may give you the wrong answer. If this is the case, then you can resort to nonparametric and contingency-table analysis.

So to use the most powerful analytic tools available, you must ensure that all assumptions—including the all-important distributional assumptions—will be met. How can you do this? Two strategies are helpful: use instruments that yield data that are continuous (not categorical), and select outcomes that are normally distributed. Many of the strategies offered in Chapter 7 for improving the quality of your measures will ensure that your data meet these criteria. For example, totaling several items rather than using a single item to measure an outcome will increase the chances that your data will be continuous and normally distributed. Your choice of measures can therefore have a big effect on statistical power.

EXAMPLE: ***Increasing statistical power by using covariates and parametric tests: The effect of a university rape-prevention program.***

Recent increases in reports of sexual assault on the nation's campuses have led some schools to initiate rape-prevention and awareness programs. Lynn Borden, Sharon Karr, and A. Toy Caldwell-Colbert (1988) investigated the effectiveness of one such program using 50 male and 50 female undergraduates at Emporia State University in Kansas. Following a pretest administration of two standardized instruments—an Attitudes Toward Rape Questionnaire and the Rape Empathy Scale—half the men and half the women in the sample participated in a 45-minute seminar on rape awareness and prevention. Although students were not randomly assigned to treatment and control groups, assignment was made on the basis of class-section meeting times, a factor the authors viewed as unrelated to attitudes toward rape. A follow-up posttest was given to both groups four weeks later.

The authors were unable to find that the rape prevention seminar had any statistically significant effect on students' attitudes toward rape. But this may not mean that the program is ineffective. As shown in Table 8.2, a total sample size of 100 provides power of only .70 to detect medium-sized effects. It may be that the rape-prevention program is actually modestly helpful, but that the sample size simply did not give sufficient statistical power to detect a true positive effect. Lack of statisical power always looms large as a possible explanation for null findings.

Nevertheless, the researchers adopted two excellent strategies to increase the statistical power of their study. First, they incorporated covariate information into their analyses—the students' gender, pattern of church attendance, and personal acquaintance with a rape victim. Second, they used sophisticated parametric analysis (multivariate repeated measures analysis of variance), a statistical procedure that was appropriate because of the high quality of the instruments they used. Even though both of these strategies increased the study's statistical power, the researchers still could not find a statistically significant effect for the program.

Borden, Karr, and Caldwell-Colbert were surprised by their null findings.

They concluded: "The nonsignificant results for the program were not anticipated because the university rape prevention program has received strong support and praise by students, as well as faculty. Indeed, there has been a steady request for the program throughout the community, indicating that it was successful in conciousness raising . . . More applied research on college campuses is needed" (p. 135). When anecdotal evidence conflicts with findings from a study, the study can override the anecdotal evidence only if it is carefully designed and has high enough power.

Instrument Precision and Sample Size

The ballpark estimates of sample size presented so far assume that your instruments are free of measurement error. But, as we discussed in Chapter 7, this is rarely the case. If your instruments have some error, you will have less power than you think. You will be less likely to detect effects that really exist, regardless of their size and your analytic technique. So before choosing a final sample size, you must consider the possibility of measurement error.

Probably the best approach to dealing with the effects of measurement fallibility on sample size is to try and improve your measures so much that you need not bother making any adjustments at all. Design away as much error as you can. The time spent improving your instruments before using them is time well spent. *Precision of your instrument is a controllable cost factor.* Don't try to save money by collecting data using less time-consuming, but less precise, instruments. Although *per-person* data collection may be cheaper, *total* data collection usually ends up being more expensive because you must collect data for more people to compensate for the imprecision of the instrument. Otherwise you sacrifice statistical power, and the savings are illusory.

Nevertheless, despite your best efforts, some measurement error may persist. If you suspect this will happen—and experience shows that it usually does—be sure to increase your target sample size accordingly. The sample sizes given in Table 8.2 are for studies that use perfectly reliable instruments (reliability = 1.00). Table 8.4 presents target sample sizes for studies that use fallible instruments, with real-world reliabilities of .60 and .80. To find a target sample size for another reliability value, simply interpolate between the two sets of numbers.

Comparison of parallel entries in Tables 8.2 and 8.4 shows the advantage of using precise measurements. As reliability decreases, your sample size must rise dramatically to ensure the same level of statistical power. For example, if your outcome is perfectly reliable (Table 8.2), you need only 113 students to have a 90 percent chance of detecting a medium correlation (.30) between it and any predictor. But if the reliability of your outcome is .80 (Table 8.4), you must study an additional 27 students to have the

TABLE 8.4. HOW MANY STUDENTS SHOULD YOU SELECT WHEN MEASUREMENT IS NOT PERFECTLY RELIABLE? BALLPARK ESTIMATES OF SAMPLE SIZES YOU NEED, ADJUSTED FOR MEASUREMENT FALLIBILITY.

		Anticipated effect size					
		Small: (reliability =)		Medium: (reliability =)		Large: (reliability =)	
Statistical test used	Statistical power	.60	.80	.60	.80	.60	.80
Pearson correlation	.90	1,741	1,306	189	140	65	47
	.80	1,302	977	142	106	49	36
	.70	1,025	769	112	83	39	29
Two-group t-test	.90	1,754	1,316	282	212	112	84
	.80	1,312	984	212	160	84	64
	.70	1,032	774	166	126	66	50

Note: Assuming a two-tailed test, alpha = .05.

same amount of power. If its reliability is .60, you must include yet another 49.

Notice that the effect of measurement fallibility on power and on sample size is most dramatic when you are looking for small effects. If you wanted a 90 percent chance of detecting a small correlation of .20, for example, the necessary increases in sample size (over what you would need if your measures were perfectly reliable) are 259 and 694 for reliabilities of .80 and .60, respectively. The bottom line: *measurement imprecision exacts a very high toll.* Try and eliminate all the error that you can.

What If Students Drop Out?

Not everyone you select for your study will agree to participate. Not everyone who agrees to participate will follow through on this intention. Not everyone who begins to participate will persevere until the end of the study. Some students drop out or are dismissed, others transfer, and many may simply forget to show up for testing and interviews. Faculty members and administrators change jobs or aren't on campus on a certain day. From the standpoint of statistical power, the reason for refusal and attrition is not important, but the disappearance of people from your sample is.

Don't be tempted to select an initial sample size just large enough to provide a specific amount of statistical power. It is not the initial sample size that counts, but the final one. You must incorporate realistic rates of refusal and attrition into your calculations of sample size.

The estimates given in Tables 8.2, 8.3, and 8.4 are the sample sizes you need to have in your *final* analyses. Because of attrition and refusal, you must increase your initial sample size to compensate for people who will disappear from your sample before analysis. If roughly 10 percent refusal and 40 percent attrition are likely, for example, you should double your initial sample size.

What rates of refusal and attrition should you expect? No single rule of thumb is particularly helpful because, even among similar studies, these rates differ widely. Some researchers have been very successful in limiting refusal and attrition. For example, in a longitudinal study of withdrawals from the University of California at Berkeley in the class of 1974, Carl Simpson and his colleagues (1980) obtained an initial response rate of 92 percent in November 1971 and a follow-up rate of 80 percent almost two years later in June 1973.

But others have not been so lucky. In a study of influences on academic growth among students at a large public university in the Northeast, Patrick Terenzini and Thomas Wright (1987) got an initial response rate of 50 percent of the 1980 entering class. On follow-up at the end of each of the four subsequent academic years, this sample fell by about 35 percent per year. By the end of the study, only 19 percent of the original sample remained.

We suggest that you make an educated guess based upon the experiences and advice of colleagues. Look for similar studies and examine their rates of refusal and attrition. Model your follow-up procedures on studies that got high rates of cooperation. Ask your registrar and personnel officers what they think you will find. Consult the admissions and student records offices. Check how many students transfer into, and out of, your school each year. Check how many students are accepted into each program and how many drop out before graduation. Check the Year Abroad programs. Consult employment and financial services offices to determine the transience of faculty and staff. If in doubt, err on the conservative side, assuming slightly *more* refusal and attrition than you really expect. After all, you can always choose to not follow up all participants, but you cannot so easily add new students to your sample once your study has begun.

EXAMPLE: *How much attrition should you anticipate: What have other researchers found?*

Common sense suggests that the longer your study, the more attrition you should expect. If you design a study with a four-year postgraduate follow-up, anticipate sizable attrition rates. Many graduates will move, others will lose contact with the alumni office, and some will not return your questionnaire. If you design a study that can be completed within a single semester, you can reduce attrition dramatically.

Many researchers have been successful at limiting attrition. Table 8.5 gives the percentage of students successfully followed over time in 10 studies we described elsewhere in this book. Not surprisingly, researchers who use short follow-up periods are particularly successful at maintaining contact with students. For example, in their one-semester studies of aca-

TABLE 8.5. HOW HARD IS IT TO KEEP ATTRITION LOW IN LONGITUDINAL STUDIES? FOLLOW-UP RATES IN TEN LONGITUDINAL STUDIES.

Author	Length of follow-up	% successfully contacted
Abrams and Jernigan (1984)	1 semester	96
Andrews (1981)	1 semester	93
Muehlenhard, Baldwin, Bourg, and Piper (1988)	4 months	87
Landward and Hepworth (1984)	1 quarter	96
	2 quarters	54
	3 quarters	50
Pascarella, Terenzini, and Wolfe (1986)	1.5 semesters	53
Simpson, Baker, and Mellinger (1980)	1 month	92
	2 years	80
Theophilides, Terenzini, and Lorang (1984)	1.5 semesters	35
	2 years	27
Terenzini and Wright (1987)	1 year	65
	2 years	42
	3 years	27
	4 years	19
Stuart (1985)	2 years	76
Hendel (1985)	5.5 years	67

demic programs, both Abrams and Jernigan (1984) and Andrews (1981) were able to retain over 90 percent of the respondents in their original samples.

Some researchers have been successful at limiting attrition even when following students for longer periods of time. After two years, for example, Simpson, Baker, and Mellinger (1980) maintained an 80 percent success rate, and Stuart (1985) maintained a 76 percent success rate. And after 5.5 years, Hendel (1985) succeeded in contacting 67 percent of his original sample, even though many of the students had graduated and left the state.

Table 8.5 illustrates that students *can* be followed over long periods of time. But this can take a real effort—many respected investigators have been unsuccessful at keeping attrition low. Rather than base your sample-size estimate on an unrealistically optimistic follow-up rate, use a conservative plan and work hard to be pleasantly surprised.