2

Defensible Program Evaluations

Four Types of Validity

DEFINING DEFENSIBILITY

Program evaluation is not someone's personal opinion about a program or someone's casual observations about a program or even observations based on journalistic or managerial familiarity with the program. Rather, it is based on defensible observations. Defensible observations are those collected in a systematic manner. "Systematic" means that the process by which the observations are collected and analyzed is both replicable and valid.

Replicability means that it does not matter who is doing the research. The basic design is laid out before the research is done. Robert Right or Leslie Left could implement the study design, collect, and analyze the data as prescribed in the study design and come to the same conclusion using the normative criteria also specified in the study design. Most of the research designs that this book discusses are *ex ante* replicable. That is, in advance, the researcher specifies in considerable detail what is to be observed, how it is to be measured, and how the information is to be analyzed. The process is known before the research is begun, but the conclusion is not. While virtually no research is 100 percent *ex ante* replicable, most research designs are quite detailed in specifying how the project is to proceed. Nonetheless, it is common for the researcher to make changes when he is in the field. He may find better ways of measuring some variables and discover that other ways are not possible. He then indicates how the design was modified, so that the audience (decision makers and others) can see, *ex post*, how the design could be replicated. Case study designs, which I briefly mention in Chapter 5 as a type of quasi experiment, have relatively more *ex post* than *ex ante* replicability. By contrast, experimental designs will have relatively more *ex ante* replicability. *Ex post* replicability means that, once the research was done, how it was done (the route that got the researcher from design to actual findings) is clear. *Ex ante* replicability means that, before the research is done, how it will be done (the route that will get the researcher from the design to actual findings) is clear.

Replicability, especially *ex ante* replicability, helps to make empirical claims (whether descriptive or causal) more defensible and objective. Without replicability (and other properties of defensibility), the claim would merely reflect personal opinion and casual observation. Replicability makes conclusions (whether descriptive or causal) traceable. That is, given the process for selecting observations and variables to study, for measuring each variable, and for analyzing the data, one can trace the link between the process and the conclusions. Traceability (or replicability) is at the core of making research objective.

While traceability (or replicability) is necessary for defensible program evaluation, it is not sufficient. If it were, one could defensibly implement a fully outlined research design that uses a process for collecting data that is known to be error prone. Such a design would be objective and replicable. It would not be valid. Validity is the second component of defensible program evaluation.

TYPES OF VALIDITY: DEFINITIONS

The four types of validity are internal validity, external validity, measurement validity (and reliability), and statistical validity.[1] I first define what these terms mean. I then discuss each type of validity in more detail, noting the threats to each type of validity and how to reduce them. The core of this text is on impact, or causal, evaluations. Because this is the province of internal validity, the discussion of internal validity begins in the next chapter. Subsequent chapters further elaborate that topic. By contrast, the discussion of external validity, measurement reliability and validity, and statistical validity will be relatively brief and is generally confined to this chapter. However, I also discuss these types of validity in more detail in forthcoming chapters when they also relate to internal validity.[2]

Internal validity refers to the accuracy of causal claims. Consequently, this type of validity is relevant only when causal evaluations are at issue. Because this is a textbook about causal, or impact, evaluations, it is a textbook about internal validity. For example, suppose that an evaluation study claimed that public school teachers with master's degrees are "more effective" than teachers with only a bachelor's degree. That is, based on the analysis of the observations in the study, the study's causal claim is that, compared with teachers with less education, teachers' additional education (including or beyond a master's degree) "causes" better performance by the students they teach. Suppose further that a subsequent, internally more valid study showed that that claim was entirely inaccurate or partially so, in that it overestimated the impact of public school teachers' advanced degrees on their students' performance. The implication would be that the internal validity of the former study was questionable. Subsequent chapters consider how to assess internal validity. It is important to note here that no study can be 100 percent internally valid. However, some studies are clearly more internally valid, or unbiased, than others. (These two expressions mean the same thing.) Reasonable levels of internal validity are necessary for causal inference if the causal claims of an evaluation are to be credible or unbiased.

External validity refers to the generalizability of research results. Research results are generalizable if they are applicable to other times and places and to the larger population that is of interest. External validity is relevant to both descriptive and causal evaluations. For example, someone interested in generalizing about state-run Temporary Assistance for Needy Families (TANF) programs in the United States would probably question the external validity of a single study of TANF programs in Nebraska, whether descriptive or causal. However, it is important to point out that one applies the criterion of external validity only to the population of interest. If one is only interested in generalizing about TANF programs in Nebraska, then the study above might be externally valid. It would not be externally valid if the population of interest were TANF programs in the United States as a whole. Temporal generalizability is another component of external validity; it is almost always relevant. That is, evaluators hope that findings from a study of ongoing programs (in a given location) will also be applicable to programs (in the same location) in the near future. Hence, findings from studies of alleged gender bias in, say, hiring practices in 1970 would probably not be generalizable to the same topic thirty-five years later, no matter how internally valid the studies were. It is also important to point out that external validity (that

is, generalizability) is sometimes not possible. If program results are contingent on the particular subgroup or location in which the program is implemented, external validity is impossible. Worse, ignoring contingency would also make the findings internally invalid if the findings represented a causal claim.

Measurement validity and reliability pertain to the appropriate measurement of all the concepts and variables in the research. Measurement validity concerns the accuracy with which concepts are measured, while reliability pertains to the precision of measurement. Another way of describing the difference between measurement validity and reliability is that valid measures have as little systematic or nonrandom measurement error as possible, while reliable measures have as little random measurement error as possible. I will discuss each of these in more detail. It is important to note that measurement reliability and validity refer not just to outcome or output variables but also to program variables and other variables that the evaluation measures.

Finally, statistical validity refers to the accuracy with which random effects are separated from systematic effects. Measurement reliability pertains to the relative absence of random error in the measure of a single variable. Statistical validity usually pertains to the relative absence of random error from the causal or descriptive claim that there is a systematic relation between variables. Thus, statistical validity usually (but not always) pertains to relations between two (or more) variables; measurement reliability usually pertains to the measurement of variables considered one at a time.

TYPES OF VALIDITY: THREATS AND SIMPLE REMEDIES

With one exception, I discuss each type of validity in greater detail. I reserve all of Chapter 3 to discussion of threats to internal validity, and Chapters 4, 5, and 6 to research designs that, to a greater or lesser extent, minimize those threats to internal validity.

External Validity

External validity refers to the generalizability of research results—that is, their applicability or portability to other times and places. Threats to external validity mainly come from four sources.

1. *Selecting a sample that is not representative of the population of interest.* If the population of interest is large (say, over 500 units of analysis), it may not be feasible or even reasonable from the perspective of costs, time, or ease of administration to study the entire population. So whom should a researcher study if she cannot study every unit in the population? The usual remedy is to select a random sample. Random means that every unit to be selected for study has a *known* probability of being selected. Many samples are simple random samples, where every unit has the *same* probability of being selected for study. More detailed texts on sampling also discuss other kinds of random samples where the probabilities are known but not equal. For example, in stratified samples, small populations may be oversampled to improve the efficiency of estimating population characteristics.[3] For the purposes of this text on program evaluation, there are three important points to make about representative sampling for external validity.

1a. *The unit of analysis is what is sampled and what is counted.* Units of analysis may be people, but often they are not. Sometimes a researcher will select a sample of individual persons from the adult population of a country or from the population of a medium or large city or from the population of students in a medium or large school district. If she selects 2,000 individuals from the population of members of a group whose population numbers about 200,000, she might select them such that each member has .01 probability of being selected. The number of observations

in the study would be 2,000. Sometimes the unit of analysis is a collection of people. Suppose a researcher wants to study schools in Maryland. There are about 7,500 schools in Maryland, and he might decide that he cannot study all of them. Instead, he decides on a representative sample, with each school having a .10 probability of being selected for study. He would collect data on 750 schools. The number of observations in his study is 750 schools; the number of observations in the study is *not* the total number of students in the 750 schools.

Similarly, suppose a researcher wants to study the implementation of Title IX programs in universities in the United States. She decides to study a representative random sample, rather than the whole population. Selecting 100 universities from a list of the 1,000 largest universities, she would design a sampling procedure such that every university in that list has a .10 chance of being in the sample. The important point here is that the relevant sample size is 100, not the number of students in the athletic programs in those universities. We revisit the issue of the units of analysis when we talk about experimental designs in Chapter 4. While the application is different, the underlying principle is the same: In selecting units for study, the number of observations in the study is the number of units in the sample, not the number of entities (if any) within the units.[4]

1b. *Small random samples are never representative.* "Small" usually means less than 120, but this is not an absolute rule. However, random samples of, say, thirty are not likely to be representative. What makes random samples representative is their size: Large random samples from a fixed population are more representative than smaller random samples. A 100 percent sample is the population, so it is clearly representative. But a random sample of 10 children is probably not representative, regardless of whether the population from which they are being selected is 300, 3,000, 300,000, or 30 million children.

Sometimes, usually because of budget and time limitations, it is not possible to study a large sample. In that case, it may be best to decide *ex ante* what a typical sample might look like. When samples must be small, rather than rely on the laws of statistical sampling that require a large sample, it is better to deliberately select what appear to be representative units from the population of interest. Such a sample is deliberately rather than randomly representative. For example, in an evaluation of a federally funded Department of the Interior (DOI) initiative to rehabilitate urban parks, the evaluators had funds to collect detailed, on-site information from no more than twenty cities.[5] Rather than select a random sample, the researchers deliberately selected cities from each region of the country; they selected cities of the size that typified those in the program, including some large cities and some small ones, and they studied different types of programs. Although a few programs provided services, most were aimed at improving the physical infrastructure of the parks (e.g., fixing the basketball courts). Consequently, the researchers deliberately selected for study more infrastructure than service programs. Had they selected twenty cities randomly, it is likely that no cities with service-oriented programs would have shown up in the simple random sample, even though they were (a small) part of the DOI initiative under study.

1c. *Random sampling from a small population will not be representative.* This is true for a variety of reasons. First, if the population is small, the random sample will also be small, and, as we have just seen, small random samples will probably not be representative of the population from which they are selected. Second, the usual theory of sampling assumes sampling with replacement. That is, in sampling theory, researchers select one unit at random from a population of, say, 10,000 units, pretend to return that unit to the population, and resample the second unit. They keep doing this until they have their desired sample of, say, 200 units. But, in practice, they do not really return the units to the population. So, the probability of selecting the first unit is 1/10,000 = .0001. The probability of selecting the second unit is 1/9,999, which is only slightly greater than .0001. The probability of selecting the third unit is also only slightly greater than the former probability. In

general, when the denominator is large, sampling without replacement (which is what is usually done in practice) is virtually the same as sampling with replacement.

However, selecting a simple random sample of states from the population of fifty states will not be random because the first unit to be selected has a lower probability of selection than the second, which has a lower probability than the third, and so on. If a researcher aims for, say, a sample of thirty, the probability of selecting the first unit is 1/50 = .02; the probability of selecting the second is 1/49 = .0204; the probability of selecting the third is 1/48 = .0208. Every unit in the analysis would have to be adjusted by the probability of showing up in the sample, adding an additional level of complexity to the analysis of the observations in the study. And the final sample may not be representative anyway because it is too small.

2. *Studying "sensitized" units of analysis.* When the units of analysis are individual people who know that they are being studied, their awareness often distorts their behavior, such that the behavior or response in the study is not generalizable to what would be observed in the real world. (Later we see that this is the same as a testing effect, a source of measurement invalidity.) For example, if bank loan officers are told that they are being studied to determine whether they service Federal Housing Administration–guaranteed mortgages differently from their own bank's mortgages, they may well behave differently in the study than they would ordinarily. Teachers who are being observed for a study may also alter their behavior, such that what is observed during a study is not representative of their ordinary behavior. The problem of studying sensitized units of analysis is often called the Hawthorne effect, based on the unexpected 1920s findings from a Hawthorne company plant that manufactured shirts. The plant managers surveyed the workers on the assembly line to see what their needs were; for example, they asked the workers whether they wanted more light to do their work. Surprisingly, the workers' output improved just after the survey, even though the managers had not changed anything. Apparently, the workers worked harder simply because the survey itself changed their behavior, signaling that management "cared."[6]

It would seem that the remedy for the problem of studying sensitized units is straightforward: Do not tell people that they are being studied. Although the respondents to a survey will be aware that they are being studied, the bank officers in our example simply need not be told that they are being studied. Similarly, social service recipients, or other program clients, simply need not be told that they are being studied. The problem with this solution is that, in general, it is illegal and unethical to fail to get informed consent from people whose behavior is being studied in an evaluation of public program implementation or impact. Although there are some exceptions to this rule,[7] the presumption is that informed consent is necessary.

An alternative strategy is to design the study so that it does not rely entirely on reactive data. Although surveys and direct observation are wonderful sources of information, they are obtrusive, and respondents may consequently alter their behavior such that it is not representative of what would be observed outside a study situation. But there are other sources of information. For example, administrative records are a source of information about the activities of teachers and bank officers in two examples that I have used. To reduce the threat to external validity from relying entirely on sensitized units of analysis, one option is to supplement the sensitive data with unobtrusive data on the same units of analysis. If the two sources of information produce similar results, then researchers can be more confident that the reactive data sources are as externally valid as the unobtrusive sources of information.

3. *Studying volunteers or survey respondents.* People who are willing to be studied may not be representative of the intended population. For example, one of the biggest problems in contemporary opinion polling and survey research is the problem of nonresponse. In this case, researchers select a large random sample of people from a specified population and phone them, send them a

survey by mail or e-mail, or visit their homes for a face-to-face interview. Although the researcher selects the intended sample, the respondents select themselves into the actual sample; in effect, they are volunteers. Typical response rates are 70 percent or lower. Even the response rate to the 2010 national census (which is not a sample; it is a tally of observations from the entire population) is only about 75 percent, and it varied considerably across the country. (For example, among states, it varied from about 65 percent in New Mexico to 82 percent in Minnesota.)

Responders are not like the population; they tend to be more educated, wealthier, and generally cooperative people. Depending on the purpose of the study or the nature of the intended sample of respondents in the study, the actual responders might be the ones with the most extreme views or more time on their hands (e.g., retired people). Another class of volunteers participates in many medical studies that compare the effectiveness of a new drug to the current drug or to a control. For example, the National Institutes of Health (NIH) offers summer internships in Washington, DC, to healthy college biology majors to work in the labs with NIH research scientists and take part in controlled drug studies. These volunteers may not be representative of the population to which the researchers would like to generalize. And, of course, many people remember being "volunteered" to be in a study in a sophomore psychology or economics class. Most people would not characterize their behavior then as representative.

Remedies for the problem of studying volunteers will only minimize the problem, not eliminate it. Chapter 7 discusses in considerable detail the steps that researchers can take to increase response rates to surveys, and I will not repeat that discussion here. The problem of generalizing from those who consent to be studied (e.g., school districts that volunteer to be in a study of school integration; college students who volunteer to be in a psychology or medical study) is usually minimized by replicating the studies in other volunteer groups. That is, if similar studies of college students from large universities, small colleges, public universities, expensive not-for-profit colleges, private universities, and the like produce the same results, the implication is that the individual studies are representative. When researchers reasonably expect that nearly all individuals respond similarly to environmental treatments, or stimuli, generalizing from volunteers or single-site studies may be valid. For example, most patients react the same way to common antibiotics, and most consumers react the same way to prices: When prices go up, people buy less. The problem of studying volunteers or sites selected by the researcher because of their convenience or availability is much more of a threat to external validity when the researcher anticipates that reactions may be different for different groups of people. This is the problem of statistical interaction.

4. *Statistical interaction.* Statistical interaction means that the descriptive relation between two variables X and Y (or the causal impact of X, the program, on the outcome Y) depends on the level or value of a third variable, Z. For example, consider a possible causal relation between public school spending and pupil achievement. Suppose that the impact of additional spending (X) on student achievement (Y) depends on the socioeconomic status (SES) of students in the school district (Z), so that more spending (X) appears to bring about ("cause") higher achievement (Y) only in low SES districts (Z–) and has no impact in high SES districts (Z+). This would be an example of statistical interaction, because spending "works" only in low-income districts. Thus, the impact of spending (X) on achievement (Y) depends on the level of district SES (Z). Similarly, if job training (X) appears effective at raising the earnings (Y) of unskilled adult women (Z_w) but not for unskilled adult men (Z_m), that also would be an example of statistical interaction.

Statistical interaction is a threat to external validity because it means that generalization is not possible. Rather, what characterizes one subgroup in the population of interest does not characterize other subgroups. When a researcher is evaluating the plausibility of causal hypotheses or causal claims, failing to recognize statistical interaction when it is present means that external

validity is not possible and reduces internal validity. That is, undetected statistical interaction can lead researchers either to erroneously find a causal relation or to erroneously reject a causal claim. Hence, we discuss the issue further in our consideration of internal validity.

The possibility of statistical interaction may also necessitate larger sample sizes to minimize the threat of small samples to external validity (and to statistical validity, as we will soon see). For instance, African-Americans comprise a small proportion of the U.S. population. If a researcher expects that a program might operate differently for African-Americans than for other groups, she might want to oversample African-Americans to ensure that there are enough African-Americans for externally valid results for that subgroup. If she is studying whether school vouchers improve academic performance among low-income public school students, anticipating that the effects might be different for black students than for white students, she should oversample African-Americans to examine this possibility. Otherwise, if the sample of African-Americans is too small, then the final causal claim about vouchers (whether the claim is "vouchers improve academic performance" or "vouchers do not affect performance") might be externally valid for the larger subgroup (those who are not African-American), but it will be less valid externally (and statistically) for the smaller subgroup of African-Americans. In fact, researchers frequently oversample many subgroups for special study simply because they anticipate statistical interaction. That is, they anticipate that it will not be possible to make one generalization about the population of interest and that the study may find that what "works" or is effective for one subgroup is not so for another.

Statistical Validity

Definition

In making descriptive or causal claims about the relation between variables (or in making descriptive claims about single variables), researchers (and critics) often wonder whether what the observations seem to show is "real" or just a fluke. For example, in the case of a single variable, if a researcher observes that achievement scores in a particular school appear extremely low compared with some external standard, that observation might be a fluke. That is, the researcher might ask, "If I did this study again (say, next week), would I get the same result? Or is the observed score just a random occurrence?" And, in the case of, say, two variables, if the researcher observed that schools with large class sizes have low achievement scores, he might ask, "Is this result real?" or "If I did this study again, would I see the same thing?" (These questions apply to both descriptive and causal claims.) Sometimes what researchers observe is purely random occurrence, especially when the number of observations (i.e., units of analysis) is small. Generalizations based on small samples are prone to random error. For example, it is quite likely that a coin, tossed twice, will show two heads (25 percent), even though it is really a fair coin. More tosses make a more statistically valid test of the hypothesis that it is really a fair coin. Random error decreases as the sample size increases.

More generally, statistical validity refers to the accuracy with which random claims (descriptive or causal) about observations are separated from systematic claims. For example, a random claim might be: "The school performance is just below the standard, but the difference is so small that it is just random." A systematic claim might be: "This school is clearly below (or above) the standard." How can we assess the accuracy of either claim? Alternatively, a random claim may pertain to the accuracy (or, in this case, precision) of a random sample: "53 percent report that they support my candidate, so it looks like my candidate may lose; the difference between winning (50 percent + 1) and 53 percent is just random." Someone else might use the same claim as systematic evidence

that the candidate will win. Which claim is more likely to be correct? Assessing statistical validity helps us to evaluate the relative accuracy, or precision, of claims like these.

Sources

Observational studies, whether they are descriptive or causal, have three sources of randomness: sampling, measurement, and human behavior. Consider, first, sampling as a source of random error. Recall that we have already related random sampling to external validity. Specifically, we noted that small random samples are likely to be low in external validity because they may not be representative of the larger population to which the evaluator wishes to generalize. Small random samples also have more random error (called sampling error) than larger samples, and thus they are more subject to problems of statistical invalidity. Statistics texts point this out, and it is not necessary to repeat those lessons here.[8] Although the probability of accuracy increases as the sample size increases, this is true only up to a point. As the sample size becomes exceedingly large (e.g., over 1,000), the probability of accuracy does not go up much, but the costs of the larger sample continue to rise, often at an increasing rate. As a consequence, we rarely observe samples of the U.S. population (or any other sample) that are much larger than 1,000.

The exception to this rule occurs when the evaluator anticipates statistical interaction. In other words, if the evaluator anticipates that, say, the impact of providing a housing voucher on housing consumption may be different for seniors than for others, such that generalization to a single population would be erroneous, then taking two separate, large samples (say, close to 1,000) of each group would increase the statistical validity of conclusions for each subgroup. The important point is that larger samples have less sampling error than smaller ones. Large samples reduce the chance that one will mistake a randomly occurring observation (noise) for a signal that is really there. Of course, larger samples always have higher costs, so researchers must balance the gain in statistical validity against the added monetary costs to determine the optimal sample size.

The ideal sample size also depends on the use that is to be made of the data. For example, we have just seen that if a researcher anticipates statistical interaction, then the ideal sample size should be larger than otherwise. Similarly, if a researcher is interested solely in the estimating population mean of one variable based on sample data, she will probably need a larger sample than if she were interested in evaluating whether a particular program is having its intended impact in a particular city. In the former case, she might need, say, 1,400 randomly selected observations (assuming that there are no issues of likely statistical interaction) in order to be 95 percent confident that an estimated mean is within + or –3 percent of the (unknown) true population mean. In the latter case, she could readily work with, say, only about 120 observations in order to be 95 percent confident that an estimate of program impact (the causal relation between X and Y, given, say, 110 degrees of freedom) is significantly greater in the intended direction than no impact at all. Further, in this case, the 120 observations could be randomly selected from the relevant population, or they could comprise the entire population of the relevant study group. Finally, in this case of impact estimation, while 1,000 observations might be better for statistical validity, it might not necessarily be optimal because of rapidly rising data collection costs. The point is that, for statistical validity, generalizing about a population's value on separate, *single* variables requires larger samples than estimating parameters that characterize causal (or even descriptive) relations *between* variables. Generalizing about the population value of single variables (e.g., mean education and median income) is usually a task for descriptive program evaluation. For statistical (and external) validity, these evaluations may require a larger n than causal evaluations.

It is also important to note that reconsidering the unit of analysis can transform what appears at the outset to be an inherently small sample with an *n* of 1 into a larger sample, simultaneously enhancing both external and statistical validity. For example, suppose that the task is to evaluate a specific shelter program that serves homeless women in a particular city. The intention is that the evaluation be generalizable only to that program. This appears to be a case in which the number of observations can only be one which is far too low for statistical or external validity.

But, in fact, this is not the case. The researcher can readily amplify the number of observations by studying a single unit over time. For example, if the shelter program has been operating for 10 years, then the potential *n* is 10 years × 12 months in a year = 120. Alternatively, and even better if it is feasible, he could compare the operation of the focal shelter, using monthly data over the past 10 years ($n = 120$), with that of a different program serving homeless women in the same city, using monthly data for the same period. Now, *n* is 240. Suppose, however, that the shelter has been in operation for only one year or that only the records for the past year are readily available. The researcher cannot then study data over time, but he can observe the entities within the study unit. Suppose the shelter, during the one-year span of time, has served 120 women. Some of the women have found independent living and employment, some are still in the shelter, and some have left and returned. A study can provide descriptive information about the program inputs and outputs for these 120 women (e.g., hours of paid employment) and even begin to examine whether the use of more program inputs "causes" better outputs (or even outcomes). In any case, *n* is 120. If the researcher can collect similar data on, say, 100 homeless women in a different program in the same city, *n* now becomes 220.

The point is that what originally appeared to be a study with one observation can be reframed by extending it over time or, by looking at individuals within a single unit, examined at a microlevel, or both, simultaneously increasing both its statistical and external validity. It may also be possible to increase *n* by adding another set of observations on individuals served by a different, comparable entity, providing a comparison for the focal institution that is being evaluated.

Two remaining sources of randomness also reduce the ability to separate systematic observations or patterns from random occurrences, jeopardizing statistical validity: randomness in measurement and randomness in human behavior. Consider first the case of randomness in measurement. We have already noted that one source of randomness in measuring population values on a single variable is small sample sizes. Just as multiple observations reduce random sampling error, multiple measures reduce random measurement error, especially when what is being measured is an abstract concept.

For instance, suppose that an evaluator is trying to estimate the employment rate of people who have completed a job-training program. Realizing that an estimate based on a random sample of ten might be a fluke, the estimate would be more precise if the random sample were 300 or 1,000. This is an example of statistical (random) error due to small sample size. In the example, employment is relatively easy to measure.

But consider the case of measuring the value of observations on a single, abstract concept like educational achievement or "social adjustment." For example, because of randomness in student performance measures, a student in a study might score low on one day, but if the same test were given the next day, the student might score better (or worse). The observation might be purely random and impermanent, but maybe it is "real" and persistent. If the observation was just a fluke, extremely low scores measured on the first day will go up the next day and extremely high scores will go down; on subsequent days, individual daily scores will fluctuate around the true mean. When there is randomness in an observed variable, any single observation will be a fluke. What is really there will be revealed by repeated measures of (roughly) the same test over time. If the seemingly low score was not a fluke, it will remain low on subsequent days, still fluctuat-

ing around the true mean. Thus, in the case of a single variable, especially when it is an indicator of an abstract concept, the best way to reduce randomness in observations or scoring is to have repeated measures or observations.

Multiple indicators, measured at a single point, also help to reduce random measurement error. As an example, consider the design of standard educational achievement tests, such as the Scholastic Aptitude Test (SAT) or Graduate Record Examinations (GRE). Why are these tests so long? Asking multiple questions about the same basic concept increases the reliability of the test (with diminishing returns) by using multiple indicators. A ten-question SAT would contain a much greater random component than the current version. Similarly, a four-question final exam in a math class would be a quicker but much "noisier" measure of a student's true performance than a fifty-item final exam.

In addition, randomness in individual-level measures (the example in these paragraphs) is usually far greater than randomness in collective-level or aggregate data (for example, school-level means or percents compared with individual-level measures), but it does not disappear, especially when the concept to be measured is abstract.

I consider the problem of random measurement error in more detail next and in Chapters 6 and 7. It turns out that random error in measures of outcome or output variables is a particular threat to statistical validity. In addition, the best way to reduce random error in the measurement of abstract concepts is to have multiple indicators or repeated measures. Just as more observations reduce random error in sample sizes, repeated measures or more indicators reduce random error in the measurement of abstract concepts. With diminishing returns, multiple indicators or repeated measures (and larger samples) separate the signal (the systematic component) from the noise (the random component). That is why researchers almost never measure an abstraction like educational achievement with just one indicator. Rather, they measure its separate components (math ability, reading ability, reading comprehension, analytical ability, and so on), using multiple items to measure each one. Multiple indicators of abstract concepts reduce the randomness in the measurement of abstract concepts.

Finally, randomness in human behavior is also a threat to statistical validity. First, randomness in human behavior is one source of random measurement error, which is due not to the measurement process but to the behavior of what is measured. This is a particular problem in survey research, but it is also a problem in other measures, too. For example, sometimes a student does well on a test, and sometimes the same student does not. The student does not know why. Sometimes she just guesses an answer; that is surely random. In surveys (or classroom tests), if students are asked to respond to a question about an issue that they have not thought about before, they respond randomly.[9] We consider the implications of random responses (in tests, surveys, and other measures) in the following paragraphs and in Chapters 6 and 7 in the discussion of measurement reliability and of surveys in evaluation research. In both of these cases, however, randomness in measures attributable to randomness in human responses makes it harder for the evaluator to separate systematic observations from random ones.

The other source of randomness in human behavior is that human behavior is very complex, probably too complex for researchers ever to completely comprehend. Furthermore, in impact evaluation research (i.e., causal evaluation studies), it is not necessary for evaluators to understand all the causes of the human behavior that they are examining. For example, suppose an evaluator wishes to estimate whether and how much a job-training program "causes" recipients to move to higher-paying jobs than they were in before. The evaluator cannot hope to explain everything about the wages of everyone in her sample. She will probably chalk up the unexplainable aspects of human behavior to the "stochastic" or random component of her study.

If the stochastic component is too large, it will be more difficult to separate any systematic impact of job training on wages from random patterns. The forthcoming chapters on research

design explain how researchers can (sometimes) make use of pretest scores to reduce the random component of human behavior without having to undertake the impossible burden of explaining it. The basic idea is that the best predictor of a person's wages at a moment in time, t, is to know what that person's wages were at a previous time, $t-1$. Taking account of the predictability or temporal stability of behavior allows researchers to increase the statistical validity of estimates of relations between program inputs and outputs, whether they are intended to be descriptive or causal. And do not forget that a large sample size is also a straightforward if not always convenient way to reduce the random component of evaluation studies. (There is another aspect to the inexplicable, random element in human behavior that is a threat to internal validity. I postpone that discussion to the extensive treatment of that topic in the chapter on internal validity.)

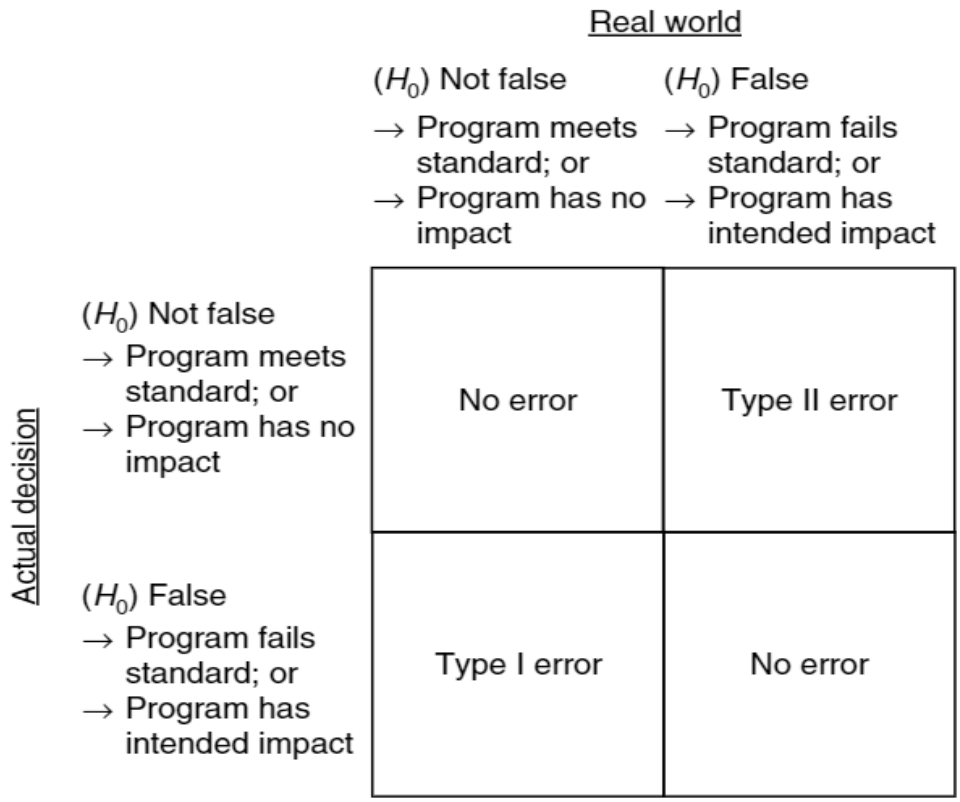
Consequences

Why is low statistical validity a problem? Low statistical validity can lead to important errors of decision. In academic research, these errors may not be costly, except to one's pride, but in program evaluation, where policy makers and program administrators must make "real" decisions based on research outcomes, these errors may well be of external consequence. No matter what its source, statistical validity tends to minimize these decision errors. In statistical language, there are two kinds of decision errors: Type I and Type II. A Type I error occurs when a null hypothesis is rejected when it is actually true; a Type II error occurs when a null hypothesis is accepted when it is actually false. Increasing sample size can reduce each type of error, but the benefit diminishes as the sample size increases. First, I characterize each type of error, and then I provide a simple illustration of how large samples can reduce the chance of each.[10]

In systematic studies, there are two kinds of hypotheses. The null hypothesis (H_0) is the one that is tested. The null hypothesis is also always an exact hypothesis. For example, in descriptive studies of a single variable, the null hypothesis might be that the observed pollution in a stream exactly meets the required (or desired) standard. In causal studies of the impact of a program on an output or outcome, a null hypothesis might be that the training program improved wages by 2 percent. Most often, the null hypothesis is that the program had absolutely no (0) impact. This (exact) null hypothesis is tested against an alternative hypothesis. The alternative hypothesis (H_1) is not exact. In evaluation research, the alternative hypothesis, while inexact, usually has a specific direction. For example, in the case of a descriptive study of a single variable, the evaluator is probably interested in whether the observed pollution in the stream exceeds the required (or desired) standard. If the pollution level is lower than the standard, no action is needed; if the pollution is above the standard, remedial action may be required. Similarly, in the case of causal evaluations, the alternative hypothesis is inexact, but it typically specifies a direction. For example, if the program manager has a standard that the training program ought to raise wages by 2 percent, impact estimates that are less than that standard may be a concern for decision makers, while beating the standard may not require action. Similarly, if the null hypothesis is that the program had no impact, the usual alternative of interest to the decision maker is that the program had an impact on the intended direction. (This is not necessarily the case in academic research, but directional alternative hypotheses are the usual case in evaluation research.)

The basic point here is that null hypotheses are exact; alternative hypotheses are inexact and usually specify a direction relative to the null. The evaluator does the study because no one knows *ex ante* which hypothesis is false. One hypothesis is false in the "real" world, but the decision maker does not know which of the two it is. The job of the evaluator is to construct a study, collect observations, and analyze the data so as to reduce the chance that the decision maker comes to the wrong conclusion. Figure 2.1 depicts the evaluator's dilemma.

Figure 2.1 **Statistical Errors in Hypothesis Tests**



The evaluator does not know which hypothesis statement characterizes the real world. Further, he can test only the null hypothesis and either reject or fail to reject it. It is not possible to "prove" a null (or an alternative) hypothesis about real-world behavior. As I remarked earlier, mathematicians do proofs; empirical social scientists do not. If the data are systematically different from what the evaluator would observe if the null were true, then he would decide that the null is (probably) not true (i.e., H_0 is false) and that the data, if they are in the same direction as the alternative hypothesis, are consistent with that hypothesis. (This does not mean that the alternative hypothesis is "true.")

Having decided that the null is not true, the evaluator may or may not be correct. Having rejected the null, he risks a Type I error, which is the error of rejecting a null that is really true. In that case, the program "really" has no impact, but the evaluator concludes (erroneously) that it does have its (intended) impact. It is also possible that the evaluator decides the null is not true and that it "really" is not true. In that case, there is no error. Alternatively, the evaluator might conclude from the study that the null hypothesis is not false. (This does not mean that the null hypothesis is true.) This might be a correct conclusion; the program may "really" have no impact. But maybe the program "really" does have an impact (in the intended direction). In this case, the evaluator has come to an erroneous conclusion. He concluded that the program had no impact, when it really does. This is a Type II error.

So no matter what the evaluator concludes, the conclusion can be wrong. Statistical validity, which is the ability to separate random from systematic influences, reduces the probability of each type of error. If the program "really" has no systematic effect (or if the sample observations are not

"really" different from the standard), then statistically valid studies will reduce the probability of erroneously finding an effect (or difference) when none is there. Similarly, if the program "really" has a systematic effect (or if the sample observations "really" are different from the standard), then statistical validity reduces the probability of erroneously finding no difference when there "really" is a difference.

No study is 100 percent valid, but some studies are more valid than others. Most important, studies with more observations nearly always reduce the probability of each type of error. However, at some point, the increase in observations begins to reduce the probability of error only a little, while the cost of collecting and analyzing more data continues to rise. In other words, at some point, increasing sample size has diminishing returns and increasing costs. So it is not the case that more observations are always better, after costs are taken into account. However, it is the case that some studies can have too few observations. A simple fable illustrates.

The Fable of the Fat Statistician

Imagine a good cookie—rich, moist, with lots of dark chocolate chips, at least two and a half inches in diameter, and one-third of an inch thick. Some cookies meet your standard of a good cookie, and others simply do not. You are a statistician; you are hungry; you are in a strange city. You go to the nearest bakery to see whether the cookies in that bakery meet your standard. You buy one cookie to take back to your hotel and test (by eating it). The null hypothesis is an exact one: The cookie meets the standard (cookie = standard). If the null hypothesis is not false, then you will buy more cookies from the bakery. The alternative (inexact) hypothesis is that the cookie fails the standard (cookie < standard). If the null hypothesis is false (meaning that the cookie appears not to meet your standard), then you will have to go elsewhere for cookies, spending more time on your search. Now you taste the one cookie, and it is OK. But based on just one cookie, you remain uncertain about whether to search further (an unwelcome thought) or to remain with this bakery and forgo a better cookie (also unwelcome). In short, you are really not sure whether to believe (i.e., fail to reject) the null hypothesis and continue to buy from that bakery or to reject the null and incur the expense and time of finding another place to buy acceptable cookies. If you reject the null but make an error in doing so, you incur unnecessary search costs and give up perfectly good cookies (Type I error). If you fail to reject the null hypothesis but make an error (Type II) in doing so, you buy more cookies that are really no good, wasting your money. So how do you reduce the chances of either kind of error? Buy (and eat!) more cookies from the test bakery. That is, you try a larger sample to increase your certainty about your decision (increase your own confidence in your judgment). That is why statisticians tend to gain weight.

The Costs of Type I and Type II Errors

In evaluation research, when policy decisions may be made based on statistical tests of null hypotheses, sometimes one type of error is worse than the other type. For example, suppose that in a political campaign, a campaign manager wants to do some research about the status of her candidate, A, against a competitor candidate, B. Her null hypothesis is that A = B, which means that the two candidates are tied in their rate of support, while her alternative hypothesis is that A > B, which means that her candidate, A, is leading. If the null hypothesis is "really" true, but the campaign manager rejects it in favor of the alternative hypothesis that her candidate is winning (a

Type I error), she may reduce her efforts to win support for her candidate when she should not. If the null is "really" false (i.e., A is actually winning), but the campaign manager accepts the null (a Type II error), then she continues to allocate excessive resources to campaigning.[11] While the Type II error is a waste of time and money, the Type I error is more costly in this case, because it may cause the candidate to lose an election.

By contrast, in impact evaluations, program managers (and program supporters) may regard Type II errors as more costly than Type I errors. For example, suppose that the null hypothesis is that a popular preschool program for poor children (e.g., Head Start) has no impact on school readiness. The alternative hypothesis is that it increases school readiness compared with what one would observe if students were not enrolled in the program. Suppose further that an evaluator, based on the data in the study, failed to reject (i.e., accepted) the null hypothesis, deciding that the program has no impact. The evaluator risks a Type II error; he also risks a storm of protest from program supporters and from the program manager. More important, if the program is canceled, but the evaluator is wrong, the children risk losing the gains in academic readiness that they would otherwise reap from what is "really" an effective program. Compared with the Type II error, the Type I error may be less costly. In that case, the evaluator erroneously rejects the null. The program continues to be funded, even though it is not effective. This too is a waste of resources, but it may not be a waste that draws as much political fire as a Type II error, at least in this case.

As another example, consider the case of the jury in the trial of a suspect who is charged with burglary or robbery. The null hypothesis is that the defendant is innocent. The jury then faces the dilemma of facing two types of errors: putting an innocent person in prison or freeing a dangerous criminal who could continue to harm society. Suppose the jury rules that the defendant is guilty, while in fact he is innocent. The jury then makes a Type I error of putting an innocent person in prison. However, suppose that the jury does not have enough evidence to reject the null hypothesis and decides that the suspect is not guilty. If he really did commit the crime, then the jury makes a Type II error, which will hurt not only the previous victims but also future ones, now that the suspect has been released. Further, the error raises the doubt that the judicial system can really punish criminals. In addition, Type II errors could eventually encourage more (severe) crimes because a few previous innocents may commit crimes hoping the justice system will let them go free. Thus on this occasion, the Type II error is arguably more costly.

In a murder case, the Type I error may be more costly than in other cases. The null hypothesis is that the suspect, charged with murder, is innocent. The jury then faces the dilemma of punishing an innocent person (perhaps with prison for life or even a death sentence), or otherwise letting a dangerous criminal go free, possibly to continue harming society. Suppose the jury does not have enough evidence to reject the null hypothesis and decides that the suspect is not guilty. If the suspect really committed the murder, then the jury would make a Type II error, which will hurt the victim's family and potential future victims. It will also fail to deter other potential killers and raise the doubt that the judicial system can really punish murderers.[12] However, if the jury rules that the defendant is guilty and sentences him to death, while in fact he is innocent, the jury then would make a Type I error. The Type I error is more costly in this case than the Type I error in the previous case of burglary or robbery. In this case of murder, the Type I error might put an innocent person to death and also raises doubts about the integrity of the judicial system.

We have seen that larger samples reduce the probability of both Type I and Type II errors.[13] Later, we will see that using "efficient" statistics and statistical tests can also reduce the probability of both Type I and II errors. Statistics like the mean (e.g., the mean SAT score in a sample of schools, which may be compared to a standard), or the difference between means (e.g., the difference between the mean standardized test scores of third-graders in comparable charter and

public schools), are point estimates based on one random sample of *n* observations. Theoretically, however, other random samples of the same size could produce other statistics. The sample from which we draw our conclusion is just one sample of many that could have been drawn. Another sample would produce a different mean (or a different difference of means). As the number of observations in our sample increases, the variance of possible means (or difference of means, or other statistical summary measures, such as a proportion) diminishes. This is desirable; we want the variation of our observed, statistical summary measure around the unknown but "true" value (that is, its likely distance or variance or standard error around the "true" value) to be smaller rather than larger. As the variation of our sample statistic around the unknown "true" value grows increasingly small, the probability of either a Type I or Type II error will go down because our guess about the "true" value based on the statistic that we actually computed from our sample data is likely to be more precise. In statistics, more precise estimates are called more efficient estimates. Thus, large samples make statistical estimators more efficient; other design aspects (including reducing random measurement error in program variables, which we discuss later) also have the same effect. Finally, we also want to estimate accurately how far our estimate is from the true value (i.e., its variance or standard error). Chapter 6 on nonexperimental designs discusses how to assess whether our estimates of the likely distance between our sample estimate and the "true" value are not only as small (or precise) as possible, but also as accurate as possible.

Alternatives to Type II Error

The issue of Type II errors is particularly vexing for program evaluators. We have already seen that program evaluators often worry more about Type II than Type I errors. For example, suppose that study results show that, of two alternatives being tested, the new alternative is no better than the current one. The evaluator fails to reject ("accepts") the null hypothesis (no difference between the treatment alternatives) relative to the (inexact) "research" hypothesis (the new program is better). But this conclusion could be erroneous because the evaluator, in deciding that the new program is no better than the old one, could be wrong. This is a Type II error. The dilemma is that, unlike null hypotheses, research hypotheses are not exact. The null is an exact hypothesis: The program had no (zero) effect. The alternative or research hypothesis is an inexact hypothesis that includes many exact hypotheses that the program had "some" particular effect in the desired direction. Given multiple sources of randomness, each of these numerous alternative exact hypotheses about program impact, even if they were "true," could produce a "zero impact" result. As a consequence, the probability of the Type II error is hard to calculate, and we do not consider that task here. There are tables of the probability of Type II errors, but the general logic is not as straightforward as that of Type I errors.[14]

However, a practical way to consider the risk of a similar error is to turn one of the alternative hypotheses in the rejection region into an exact one. For example, having decided in step 1 not to reject the null, the evaluator, in step 2, could next test the observed results against a minimum acceptable threshold of desired effectiveness. The minimum threshold becomes an exact hypothesis, and the research proceeds in the usual way. The minimum threshold could be what political decision makers (or program managers) consider minimally acceptable. That level could be outside the .05 rejection region (especially if the sample size is small), but it could still be better than "no impact" from a management point of view. The threshold could be a level determined by legislative statute or court order, or it could be the break-even point in a cost-benefit or cost-effectiveness analysis. So, having accepted the null (no effect) hypothesis test, the evaluator can next test the observed results from the study against the minimum acceptable threshold, which now becomes

the exact null hypothesis that is tested in the second stage of the analysis. While the computed *p*-value from this second stage is technically the probability of a Type I error, it also provides information about the probability of incorrectly deciding that the program does not work (i.e., it fails to meet the standard) when in fact it may work at an acceptable level.[15]

In a similar vein, Jeff Gill suggests paying attention to the confidence interval of a parameter estimate.[16] Confidence intervals decrease as the sample size increases, which is analogous to increasing the power of a null hypothesis test, in which power is the probability that failing to reject the null is the correct decision. This may be less confusing than a hypothesis test, since there is no Type II error in estimating a confidence interval.

In sum, it is particularly important in program evaluation to avoid rigid adherence to a hypothesis-testing model of the 0-null hypothesis using a conventional *p*-value of .05. In academic research, real careers may depend on statistical decisions, but in program evaluation, real programs, as well as real careers, are at stake. The best advice is to use multiple criteria. If the program is acceptable (or unacceptable) under multiple criteria of statistical validity, then the statistical decision becomes more defensible. However, statistical validity is not the only criterion for the valid assessment of program characteristics or impact. I turn next to the critical issue of measurement.

Measurement Reliability and Validity

Introduction

Valid descriptions of program inputs, outputs or outcomes, and valid assessments of program impact require that the measures of program inputs and outputs or outcomes themselves are defensible. For example, if an evaluator is examining the impact of participatory management on productivity in a school, she needs to have valid measures of management that are more or less participatory and valid measures of output that represent productivity levels.

Abstract concepts like these are particularly difficult to measure. In fact, the overall measurement of "validity" is parsed into separate criteria: reliability and validity. The reliability of a measure is the *absence* of random measurement error (RME) in recorded scores. The validity of a measure is the *absence* of *non*random measurement error (NRME) in recorded scores.

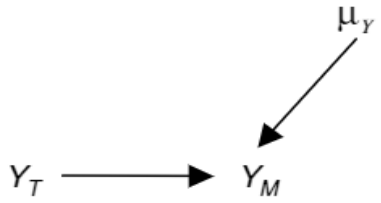
A diagram is the best way to distinguish between reliability (no random error) and validity (no nonrandom error) in measures. Consider a measure of school productivity using test score gains in a first-grade classroom. Call that measure Y. Y has two components. First, there is the "true" score Y_T; we do not know what it is. We only know what we observe or measure, which we call Y_M. Figure 2.2 shows how Y_T is related to Y_M. In this diagram, the measured scores (Y_M) are determined by a random component (μ_Y) and a systematic, or nonrandom, component (Y_T). If most of Y_M consists of μ_Y, then Y_M is a noisy measure of Y_T, with considerable RME. However, if most of Y_M is due to Y_T, then Y_M is likely to be a relatively valid measure of Y_T, with little NRME.

Representing this diagram algebraically is more informative, especially with respect to NRME. Specifically, we write Y_M as a linear function of both Y_T and μ_Y:

$$Y_M = \alpha + \beta Y_T + \mu_Y$$

In this formulation, if the expected value of μ_Y is small [written $E(\mu_Y)$] (and if it has little variance), then we would conclude that Y_M has little RME. With respect to NRME, $E(\alpha) = 0$ and $E(\beta) = 1$ (and they have little variance), then we would conclude that $Y_M \cong Y_T$, so that the measured and true scores of Y are about the same. One could have a valid measure Y_M with considerable RME:

Figure 2.2 **The Basic Measurement Model**



$Y_M = 0 + 1 * Y_T + \mu_Y$. Alternatively, one could have an invalid measure of Y_M with little RME: $Y_M = \alpha + \beta Y_T + 0$, where the intercept is not expected to be 0 and the slope is not expected to be 1.[17]

Below, I discuss examples of both kinds of measurement errors, including problems of likely RME and NRME in test scores of first-graders. To begin the discussion, consider my beloved, but old bathroom scale. In the long run, scales like these have little (but not zero) random measurement error. In fact, I can see the random error on this old analog scale. When I step on it, the indicator bounces around a little before it settles down to a number. Thus, I assume that my scale has relatively little RME: $E(\mu_Y) \gg 0$. However, my scale has considerable NRME. First, it consistently underreports my weight; symbolically, this means: $0 < E(\beta) < 1$. Second, it is anchored at a negative weight, so it registers a slightly negative score when no one is standing on it: $E(\alpha) < 0$.

While a bit lightweight, this example serves to illustrate the two aspects of measurement error (random and nonrandom). It also illustrates the two facets of NRME: constant error or error in the intercept; and systematic error or correlated error in the slope. Examples and implications of these errors for assessing overall measurement reliability and validity follow.

Measurement Reliability

Measurement reliability refers to a measurement procedure that has little RME. I have already suggested some examples of measurement procedures that are likely to contain random components:

- standardized achievement tests
- classroom tests
- athletic ability
- responses to attitude or opinion survey questions, particularly when the issue is unfamiliar to the respondent or when the respondent has ambiguous attitudes or beliefs
- nonhomicide crime rates
- safety (of anything)

Measurement reliability is achieved if different measures of the same phenomenon record the same results. For example, continuing with the example of my bathroom scale, the scale is a reliable measure of weight if I step on it and it reads 112. I step off and then step on it again two minutes later, having done nothing except maybe read a few pages of this book. Once again, the scale reads 112. I conclude (if I did such a test repeatedly, with the same or close to the same results each time) that the scale is reliable.

By contrast, we say that, compared with the bathroom scale, an SAT score is not as reliable an indicator of academic achievement because a student taking the SAT twice in a short period may get different scores, even though her underlying level of achievement remains unchanged.

Similarly, we are accustomed to getting different scores on exams or in athletic competitions, and we often attribute surprising success to "good luck" and surprising failure to "bum luck." The technical term for these casual assessments is "random measurement error."

Test scores may be particularly unreliable, but randomness is greater under some circumstances than others. At the individual level, scores on standardized tests fluctuate randomly. However, larger "samples" reduce randomness. For example, increasing the number of items on a test reduces randomness in the overall test score of an individual. At the classroom, group, or school level, randomness in the group average decreases as the number in the group increases. Thus, test scores for minority groups may contain more "noise" than scores on the same test for nonminorities. Statistically, this is recognized as a problem of heteroscedasticity in measurement. Scholars also point out that failure to recognize random error in test scores (and in measures of test score gains) is likely to result in underestimates of teacher effectiveness.[18]

Responses to opinion surveys provide another, less familiar, example of often-unrecognized unreliability. According to Asher, respondents to opinion surveys commonly feel pressured to say something when they are asked a question in a poll.[19] This reaction is particularly likely when respondents have ambivalent opinions about a complex topic (like the death penalty). It is also likely when respondents know little or nothing about the topic or if the topic is a nonsense question (e.g., "Should the music industry increase the level of hemiola in contemporary music, reduce it, or is the current level satisfactory?"). The actual response will be "random," but it will be indistinguishable from a "real" one, unless the possibility of random response is anticipated.

Even crime rates, which look "real," contain RME, because not all crime is reported, and sometimes the decision of a citizen to report a crime is just a random one. (Sometimes, the decision to report reflects characteristics of the reporter and is not just a random phenomenon; I discuss NRME later.) Homicides, however, are likely to be reported, so they are not likely to be subject to problems of random (or nonrandom) measurement error. This also characterizes accident data. For example, small accidents are not consistently reported to authorities, and some of the nonreporting is undoubtedly random (and some probably reflects characteristics of the reporter, so part of the error is not random). Significant accidents (e.g., those that result in death or hospitalization) are more likely to be reported. Thus, data on automobile fatalities are probably more reliable than data on the number of nonfatal automobile injuries.

RME may also plague what appear to be objective measures of program implementation. Sometimes, what is recorded may reflect random reporting errors or data entry errors. For example, if the evaluator is studying the impact of hours spent in a job-training program on the probability of finding a job, the reported measure of finding a job (yes or no) may be quite reliable. However, the measure of hours spent in training may not be as reliable because random errors frequently plague administrative recordkeeping, particularly when the agency providing the training is not a large bureaucracy that is accustomed to keeping records and can do so at low marginal costs.

Consequences of RME and Strategies for Reducing It

Virtually no measurement is 100 percent reliable, but some measures are more reliable than others. Why should program evaluators, concerned about making their research conclusions defensible, care about reliable measures? It turns out that unreliable measurement has one and sometimes two undesirable effects on validity. First, unreliable measures of outcome variables reduce statistical validity, thus raising the likelihood of both Type I and II errors. Chapter 6 on nonexperimental design discusses this further. Second, unreliability in measures of program

variables (but not output variables) also reduces internal validity. The next chapter, on internal validity, makes clear why internal validity is particularly important for defensible program evaluation results, and Chapter 6 explains why NRME in program variables adversely affects internal validity.

Contrary to intuition, it is more important to be concerned that program and treatment variables are measured as reliably as possible than it is to focus attention on reliably measuring outcome or output variables. Yet conventional wisdom is to concentrate on reliable measures of outputs or outcomes, but the cost may be to ignore the development and assessment of reliable measures of program treatment. In program evaluation, it is not clear that the gains in statistical validity from concentrating on reliably measuring outcomes or outputs are worth the losses in internal validity if reliable measures of program treatment are sacrificed. To give some examples, program evaluators tend to concentrate on reliably measuring outcomes or outputs (e.g., achievement scores, wages, recidivism, compliance). If the measures of program treatments (e.g., hours in school, hours in a training program, hours in a drug treatment program, quality and quantity of safety or health inspections) are unreliable, then estimates of the impact of X on outcome Y could be internally invalid even if the outcome Y is reliably measured.

A relatively straightforward way to increase the reliability of measurement is not unrelated to the way to increase statistical validity. Just as increasing the number of observations in a study reduces random error, so does increasing the number of indicators reduce measurement unreliability. I prove this statement in Chapter 7, which discusses measurement of the reliability of responses to survey questions. But the rationale for this statement is easy to demonstrate with a simple example. Suppose that your performance in this (or any other) class was to be assessed with one exam; on that exam, there is only one short-answer question. While students and the instructor would all enjoy the reduced workload, most students would complain that one short question on one exam is not a very reliable measure of their performance in the class. Maybe you will have a cold on that day. Or the question does not touch on what you have spent most of your time working on. Or the question deals with the topic that you found the hardest to understand. Or maybe you got lucky, because the test question represents the *only* thing that you understand from the course. The point is that one item measuring a student's performance in an entire class is an unreliable measure. More items, and more tests, increase reliability.

Using multiple indicators or multiple items to increase the reliability of measurement is particularly important when concepts are hard to measure (e.g., outcome or output measures such as class performance, satisfaction with a program, environmental quality, or wellness). By contrast, when concepts are not so abstract or hard to measure (like weight, hourly wages, or hours of work), multiple indicators are not as important because a single indicator can be reasonably reliable.

Frequently, evaluators combine multiple indicators into a single index. Your grade in this class is such an index, assuming that it is an average (weighted or unweighted) of grades on several tasks. The final score in a baseball game is an index of the performance of each team in an inning. Your overall SAT or GRE score is an index of your performance on each item in the test, and each component of the exam (e.g., the verbal score) is an index of your performance on each item in that portion of the test. Chapter 7, in addition to discussing how to measure reliability, also discusses how to create indexes and assess the reliability of indexes. Indexes that have more components (more indicators) are likely to be more reliable than indexes with fewer components. For example, a ten-item test is usually more reliable than a three-item test. Chapter 7 also discusses how to assess the validity of indexes.

Measurement Validity

Measurement validity is different from measurement reliability. While measurement reliability refers to a measurement procedure that is (relatively) absent of random error, measurement validity refers to a measurement procedure that is (relatively) absent of NRME. NRME means that the measure is biased—that is, it contains a nonrandom component that does not have anything to do with what the researcher really wants to measure, so the measured score Y_M is not equal to Y_T, even with allowance for RME. Recall that there are two types of NRME: bias in the intercept (constant or consistent bias); and bias in the slope (systematic bias in the measured score Y_M that is correlated with the true score Y_T). It turns out that one way to increase measurement validity is the same as the way to increase measurement reliability: Use multiple indicators.

First, let us consider some examples of possible NRME. A common charge is that SAT and GRE scores are biased. Specifically, the charge is that minorities whose true score = Y_T perform more poorly on these tests than nonminorities, so their observed score $Y_T < Y_M$. The deviation is allegedly not random; rather, it is allegedly due to the race or ethnicity of the test taker, which is not what the test is supposed to be measuring. This is an allegation of potential bias due to NRME in SAT, GRE, and other standardized, multi-item, reliable achievement test scores.

If the allegation were true, it would be an example of systematic error. Even if there is no direct connection between underlying true scores and race, racial minorities in the United States typically come from families with low financial and human capital assets. One consequence is low measured test scores, as a result of low assets, not race. Using the basic measurement model to represent this allegation, where $Y_M = \alpha + \beta Y_T + \mu_Y$, there is not only RME but $E(\beta) < 1$, unless race and capital assets (often measured by indicators of socioeconomic status [SES]) are accounted for. In this example, assume $E(\alpha) = 0$; there is no intercept or constant error. Rather, the measurement error affects the relation between Y_M and Y_T, which, unless otherwise accounted for, reflects the direct impact of SES on measured test scores. Figure 2.3 represents this dilemma. Unadjusted GREs, SATs, and other standardized test scores do not account for these alleged sources of systematic, or correlated, bias.[20]

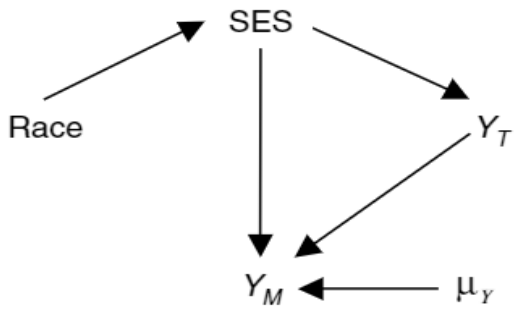
Another example of correlated or systematic NRME is race-of-interviewer effects on responses to face-to-face surveys. Apparently, respondents alter their responses to many (but not all) survey items depending on whether the interviewer is the same race as the respondent.[21]

In addition to distinguishing between constant and correlated NRME, researchers make other distinctions to characterize NRME. These distinctions overlap our distinction between constant and correlated NRME. For example, it is useful to describe three aspects of NRME in the following manner:

1. *Face validity:* Does the actual indicator reflect what it is supposed to measure? For example, students often argue that a final exam did not reflect what was taught in the class or reflected only a small component of what was taught. That is an allegation of face invalidity. I have always wondered whether scores on driving tests (the written plus the behind-the-wheel component) in the United States really indicate a driver's ability to handle a car skillfully and safely. A spelling test alone would, on its face, be an invalid indicator of a student's overall verbal ability.

These are examples not only of face invalidity but also of intercept or constant NRME: The allegation is that a high score on a typical driver's test or spelling-only test overestimates actual driving or verbal ability. Using multiple indicators (e.g., for verbal ability: a spelling test, a test of reading comprehension, and a test of the ability to compose an explanatory paragraph) would go far to improve face validity, just as it improves reliability.

Figure 2.3 **The Model of Systematic Nonrandom Measurement Error**



2. *Concept validity:* Are the measured indicators of the same concept correlated with one another and uncorrelated with unrelated concepts? (This is also called convergent and discriminant validity, respectively.) For example, if academic achievement (e.g., grade point average) is correlated with four related indicators (e.g., scores in SAT verbal, SAT math, SAT reading, and SAT reasoning), then we might regard these as valid indicators of the concept "academic achievement." However, if any (or all) of these indicators correlate with an unrelated concept, such as race, we would regard that as a sign of concept invalidity. They are also examples of systematic or correlated NRME.

3. *Predictive or criterion validity:* Do the indicators predict the expected outcome? For example, does the score on a driver's test (written plus behind-the-wheel performance) overestimate or accurately predict a person's ability to drive? Does a high GRE score underestimate or accurately predict a student's performance after she gets into graduate school? In either case, does the prediction depend also on, say, the (unobserved) personality of the test taker? These questions raise issues of predictive validity. They are also instances of both correlated and constant NRME.

Measurement Errors: Threats to Statistical or Internal Validity

Measurement reliability and validity are problems for the validity of evaluation studies for several reasons. First, in causal evaluations, RME in any variable except the output or outcome variable will reduce the internal validity of any causal claim, regardless of whether the claim is "there is an impact" or "there is no impact." NRME in any variable will also reduce the internal validity of a causal claim. Forthcoming chapters, which list and explain threats to internal validity, develop the relation between measurement issues and internal validity. Second, in both causal and descriptive evaluations, RME in variables reduces the statistical validity of the evaluation study. It is never possible to have 100 percent reliable and valid measurement procedures, but some measurement procedures are more reliable and valid than others.

In general, as we have seen, the best way to improve measurement reliability is to use multiple indicators of program treatment and program outcome. Usually, this also improves face validity and may well reduce other sources of NRME. Chapter 1 stressed the normative importance of multiple indicators of outputs and outcomes. Chapter 7 on surveys briefly introduces alpha as a measure of reliability and factor analysis as a tool for assessing the validity of measurement procedures; both are useful whenever there are multiple indicators. Multiple indicators are thus central to measuring complex concepts: Having multiple indicators allows researchers to assess both reliability and validity and also is likely to improve both. Proper model specification for internally valid estimates of program impact, considered in the next chapter, and the use of statistical controls, considered in Chapter 6, are also essential for reducing systematic (or correlated)

NRME. Because of the connection between systematic NRME, RME in program variables, and internal invalidity, separating measurement reliability and validity, the topics of this chapter, from internal validity, the topic of the next chapter, is rather artificial. Thus, it is important to turn to the more general issue of internal validity.

BASIC CONCEPTS

Defensible designs
Replicability
Internal validity: definition
External validity: definition
Statistical validity: definition
Measurement reliability: definition
Measurement validity: definition
Threats to external validity
Unrepresentative sample
Sensitized units of analysis in sample
Volunteer respondents
Statistical interaction
Threats to statistical validity
Random sampling error
    Making *n* larger
Random measurement error
    Making number of indicators larger: multiple indicators
Random human behavior
Statistical errors
    Type I
    Type II
Costs of statistical error
    Type I costs
    Type II costs
Alternatives to Type II error
Threats to measurement validity: the measurement model
Diagram: RME vs. NRME
Equation: RME vs. NRME
RME: examples
RME: consequences
    RME in program variables (X)
    RME in output/outcome variables (Y)
Reducing RME: multiple indicators
NRME: examples
Constant NRME
    Correlated/systematic NRME
    Face invalidity: constant NRME
    Concept invalidity: correlated NRME
    Predictive invalidity: constant NRME
Reducing NRME: multiple indicators

DO IT YOURSELF

Find an example of an evaluation of a public or nonprofit program, management initiative, or recent reform effort. The evaluation could concern a program in the United States or in another country. The evaluation need not be an impact evaluation. It may simply describe program performance. In the United States, most federal government agencies are required to report their performance according to Government Performance and Results Act standards, and links to that information can be found in agency Web sites. That would be a convenient source of information for this exercise. There are many published or unpublished evaluations of local government agencies, especially school districts, schools, and police departments, either by outsiders or insiders. Newspapers often report the results of these evaluations; the original evaluation is a good source for this exercise. The World Bank continuously evaluates projects that it funds, and so does the Ford Foundation; these provide another source of information for this exercise. Warning: The exercise looks simpler than it is.

The Exercise

Evaluate the "evaluation" according to the following criteria:

- External validity: How generalizable are the conclusions?
- Statistical validity: How well does the report separate the signal from the noise?
- Measurement reliability and validity:
  - How noisy are the measures? (reliability?)
  - Are the measures reasonable estimates of the "true" underlying concept? (constant measurement error)
  - Are the measures likely to be correlated with factors that do not reflect the underlying concept? What factors? (correlated measurement error)

NOTES

1. Richard A. Berk and Peter H. Rossi, *Thinking About Program Evaluation*, 2d ed. (Thousand Oaks, CA: Sage, 1999), also use this classification of the types of validity. Chapter 6 shows that the four types of validity can be reduced to just internal and statistical validity. Measurement reliability and validity (and, arguably, external validity) are important because they are aspects of internal and statistical validity.

2. Readers can refer to Leslie Kish, *Statistical Design for Research* (New York: Wiley, 2004), for external validity and sampling; Madhu Viswanathan, *Measurement Error and Research Design* (Thousand Oaks, CA: Sage, 2005), John McIver and Edward G. Carmines, *Unidimensional Scaling* (Thousand Oaks, CA: Sage, 1981), Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment* (Thousand Oaks, CA: Sage, 1979), Jae-On Kim and Charles W. Mueller, *Factor Analysis: Statistical Methods and Practical Issues* (Thousand Oaks, CA: Sage, 1978), and David Andrich, *Rasch Models for Measurement* (Thousand Oaks, CA: Sage, 1988), for measurement; and Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences*, 4th ed. (Upper Saddle River, NJ: Prentice Hall, 2009), for general statistical topics.

3. Agresti and Finlay, *Statistical Methods for the Social Sciences*.

4. Relevant to the unit of analysis is the issue of ecological inference. If the data in a study are data about schools, one cannot usually claim that the unit of analysis pertains to students. Similarly, if the data in a study are data about cities, one cannot claim that the study results pertain to individuals in the cities. For example, if one finds that high immigration rates in cities have no impact on crime in those cities, one cannot make a straightforward claim about immigrants and crime at the individual level. For further discussion of the problems of ecological inference, see Laura Langbein and Alan Lichtman, *Ecological Inference* (Beverly Hills, CA: Sage, 1978); Christopher Achen and W. Phillips Shively, *Cross-Level Inference* (Chicago: University of Chicago Press, 1995); Gary King, *A Solution to the Ecological Inference Problem* (Princeton: Princeton University Press, 1997). Some studies have multiple

levels of units of analysis. For example, there are studies of students within schools and schools within school districts. One might study immigrants within cities and cities within states. Chapter 6 includes a brief discussion of hierarchical (multilevel) modeling in evaluation research and provides references.

5. William Martineau, Laura Langbein, Richard White, and Peter Hartjens, *National Evaluation of the Urban Park and Recreation Recovery Program* (Washington, DC: Blackstone Associates, 1983).

6. Steven D. Levitt and John A. List report that much of the story of the Hawthorne effect is an urban myth. They actually located the original data, reanalyzed it using the methods described in this book, and found no lighting effect or any clear observational effect. They do not deny that subjects often react to being observed (or surveyed); they point out that the "design" in the original study was deficient. See Levitt and List, "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments," *Applied Economics* 3, no. 1 (January 2011): 224–238.

7. American Evaluation Association, *Guiding Principles for Evaluators*, 2004, www.eval.org/gptraining/GPTraining Final/gp.principles.pdf (accessed July 5, 2011).

8. For example, Agresti and Finlay, *Statistical Methods for the Social Sciences.*

9. Robert Weissberg, Policy, *Polling and Public Opinion: The Case Against Heeding the "Voice of the People"* (New York: Palgrave Macmillan, 2002); Herbert Asher, *Polling and the Public: What Every Citizen Should Know* (Washington, DC: CQ Press, 2001).

10. Mark Lipsey observes that, in a single study, there can be only one type of error. If the H_0 is really true, the only possible type of error is a Type I error. If H_0 is really false, then Type II is the only possible error. It is particularly difficult to assess the probability of a Type II error, because it depends on the (unknown) effect size parameter. Lipsey points to the role of meta-analysis (see Chapter 8) to help assess effect sizes. Based on multiple studies, such effect size estimates can be used to estimate the probability of a Type II error. See Mark W. Lipsey, "Statistical Conclusion Validity for Intervention Research: A Significant ($p < .05$) Problem," in *Validity and Social Experimentation: Donald Campbell's Legacy*, vol. 2, ed. Leonard Bickman (Thousand Oaks, CA: Sage, 2000), ch. 4.

11. In a statistics text, the campaign manager in this situation would not "accept" a null hypothesis; rather, the manager would "fail to reject" the null. In program evaluation, managers often take real actions based on statistical decisions. Thus, if they fail to reject the null, they take action based on the belief that the null is (probably) true. In effect, they act as if the null were true; they effectively "accept" the null.

12. There is controversy over the specific and general deterrence impact of capital punishment. If capital punishment has no deterrent impact, the cost of a Type II error would be considerably less. See Isaac Erlich, "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *American Economic Review* 65, no. 3 (1975): 397–417; Brian Forst, "Capital Punishment and Deterrence: Conflicting Evidence?" *Journal of Criminal Law and Criminology* 14, no. 3 (1983): 927–942; Hashem Dezhbakhsh, Paul H. Rubin, and Joanna M. Shepherd, "Does Capital Punishment Have a Deterrent Effect? New Evidence from Postmoratorium Panel Data," *American Law and Economics Review* 5, no. 2 (2003): 344–376; Robert B. Ekelund, Jr., John D. Jackson, Rand W. Wessler, and Robert Tollison, "Marginal Deterrence and Multiple Murders," *Southern Economic Journal* 72, no. 3 (2006): 521–541. Erlich's article was the first to test and support the deterrence hypothesis; it set off a plethora of empirical research, reviewed by Forst, that raised significant doubt about the deterrence effect of capital punishment. The articles by Dezhbakhsh et al. and Ekelund et al. use newer data and more recent methodological improvements to reaffirm Erlich's original findings. Chapter 6 on nonexperiments discusses the panel methods used in those articles.

13. Juries cannot conveniently, legally, or morally increase the sample size. Instead, they can reduce the probability of error by increasing the amount of evidence (not unlike increasing the sample size) for the single case that they are considering.

14. Christopher L. Aberson, *Applied Power Analysis for the Social Sciences* (New York: Routledge, 2010); Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Lawrence Erlbaum, 1988).

15. There are important problems of logic and misinterpretation that surround null-hypothesis significance testing. The most noteworthy logical challenge is posed by Bayes' theorem. The classical model, common in program evaluation (and in this text), asks the question: If the "theory" is true (e.g., the program has no impact; the program meets the standard), what is the probability of observing the results that we got from the data that we collected? Bayes' challenge inverts the question and deals more explicitly with the issue of Type II error. It asks: What is the probability that the theory is true (e.g., the program meets or beats the standard, which is an inexact hypothesis), given the results that we observe from the data we collect?

Misinterpretations are also common. For example, statistical significance in a large sample study should not be confused with real-world importance or substantive significance. Similarly, a given *p*-value in a large sample study is no more reliable (i.e., less random) than the same *p*-value from a smaller sample study. It

should also be clear from the preceding discussion in this chapter of the "real" costs of Type I (vs. Type II) error in program evaluation, relative to academic research, that Type I *p*-values of .05 have no particular magic, especially in comparison to a *p*-value of .049 or .051. Finally, failing to reject the null does not rule out any of the infinite number of competing hypotheses. See David Rindskopf, "Plausible Rival Hypotheses in Measurement, Design, and Scientific Theory," in *Research Design: Donald Campbell's Legacy,* vol. 2, ed. Leonard Bickman (Thousand Oaks, CA: Sage, 2000), ch. 1; and Jeff Gill, "The Insignificance of Null Hypothesis Significance Testing," *Political Research Quarterly* 52, no. 3 (1999): 647–674.

Stephen Ziliak and Deirdre McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* (Ann Arbor, MI: University of Michigan Press, 2008), and Deirdre McCloskey and Stephen Ziliak, "The Standard Error of Regressions," *Journal of Economic Literature* 34, no. 1 (1996): 97–114, also point to the importance of having the relevant null hypothesis: If the substantive focus is on how much consumption (e.g., of gasoline) responds to higher prices, then the relevant null is probably not that the relation is 0; rather, the relevant null is that demand is price elastic, which implies a null hypothesis of –1.00. In addition, they make the critical point that statistical significance has nothing to do with substantive significance. Results that are statistically significant may be substantively trivial; this is particularly likely with large sample sizes, which can make almost anything look significant. When sample sizes are very small, substantively important results often look insignificant. For example, if a standard of health is "close to" 3,500 units (e.g., white blood cells) per 100 milliliters, and two subsequent random measurements with a known (i.e., valid and reliable) instrument ($n = 2$) report results of 40K and 75K units per 100 milliliters, respectively, further inquiry if not remedial action should proceed, regardless of the outcome of a significance test.

But if a large n (e.g., $n > 50{,}000$) upholds a null hypothesis of "no impact" (and if other aspects of the research design are defensible), the probability of a Type II error is vanishingly small, making the "no effect" conclusion more credible. For example, see Christopher Carpenter, Sabina Postolek, and Casey Warman, "Public-Place Smoking Laws and Exposure to Environmental Tobacco Smoke (ETS)," *American Economic Journal: Economic Policy* 3, no. 3 (2011): 35–61, who find that laws banning smoking in public places (in Canada) reduce exposure to ETS but have no impact on the number of cigarettes smoked by smokers. In all their analyses $n > 65{,}000$. In this case, statistical significance is an "easy" standard to reach, and the authors point to the substantive significance of the "significant" results and also to the "significance" of the insignificant results, which with such a large n, is in fact a "hard" standard to reach.

16. Gill, "Insignificance of Null Hypothesis Significance Testing," and Ziliak and McCloskey, *The Cult of Statistical Significance*, also suggest this.

17. See Madhu Viswanathan, *Measurement Error and Research Design* (Thousand Oaks, CA: Sage, 2005), for a complete text on measurement that also distinguishes between random and nonrandom errors of measurement and categorizes nonrandom errors into two basic components: additive (i.e., constant error) and correlational (i.e., correlated) error. Viswanathan further distinguishes between within- and between-measure correlated errors. Most of our examples illustrate between-measure errors. An example of within-measure error occurs when some survey questions on political liberalism are positively worded, while others are negatively worded.

18. See Don Boyd, Pam Grossman, Hamp Lankford, Susanna Loeb, and Jim Wyckoff, "Measuring Effect Sizes: The Effect of Measurement Error" (paper presented at the National Conference on Value-Added Modeling, Madison, WI, April 2008); Eric A. Hanushek and Steven G. Rivkin, "Generalizations About Using Value-Added Measures of Teacher Quality" (paper presented at the annual meeting of the American Economic Association, Atlanta, GA, 2010).

19. Asher, *Polling and the Public*.

20. See Arthur C. Brooks, "Evaluating the Effectiveness of Nonprofit Fundraising," *Policy Studies Journal* 32, no. 1 (2004): 363–374, on the use of adjusted measures of the performance of nonprofits to neutralize the effects of uncontrollable outside forces. See also Arthur C. Brooks, "The Use and Misuse of Adjusted Performance Measures," *Journal of Policy Analysis and Management* 19, no. 2 (2002): 323–328. See also Gary King, Christopher J.L. Murray, Jashua Ajay Salomon, and A. Tandon, "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research," *American Political Science Review* 98, no. 1 (2004): 191–207, on problems of constant and correlated NRME in survey-based measures of self-assessed visual acuity (and other measures of personal health) in cross-national research and many other applications as well. The authors of the last-named article also suggest a way to adjust observed measures for this bias.

21. Darren W. Davis, "The Direction of Race of Interviewer Effects Among African-Americans: Donning the Black Mask," *American Journal of Political Science* 41, no. 1 (1997): 309–322. See also Gabriele B. Durrant, Robert M. Groves, Laura Staetsky and Fiona Steele, "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys," *Public Opinion Quarterly* 74, no. 1 (2010): 1–36.