# Public Program Evaluation

# Public Program Evaluation

## A Statistical Guide

Laura Langbein with Claire L. Felbinger

*M.E.Sharpe*
Armonk, New York
London, England

**Library of Congress Cataloging-in-Publication Data**

Langbein, Laura.
Public program evaluation : a statistical guide / by Laura Langbein with Claire Felbinger.
p. cm.
Includes bibliographical references and index.
ISBN 13: 978-0-7656-1366-0 (cloth : alk. paper)
ISBN 10: 0-7656-1366-2 (cloth : alk. paper)
1. Policy sciences—Statistical methods. 2. Evaluation research (Social action programs)—
Statistical methods. I. Felbinger, Claire L. II. Title.

HA97.L35 2006
352.3′5—dc22 2006005777

Printed in the United States of America

The paper used in this publication meets the minimum requirements of
American National Standard for Information Sciences
Permanence of Paper for Printed Library Materials,
ANSI Z 39.48-1984.

∞

EB (c) 10 9 8 7 6 5 4 3 2 1

# Contents

# 3 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 we discuss in this book 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 we briefly discuss in Chapter 6 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 A to B) 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 A to B) is clear.

Replicability, especially *ex ante* replicability, helps to make empirical claims (no matter whether they are 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 (no matter 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

There are four types of validity: internal validity, external validity, measurement validity (and reliability), and statistical validity.[1] We first define what these terms mean. We then discuss each type of validity in more detail, noting the threats to each type of validity and how to reduce these threats. The core of this text is on impact, or causal, evaluations. Because this is the province of internal validity, our discussion of internal validity begins in the next chapter. Subsequent chapters further elaborate that topic. By contrast, our discussion of external validity, measurement reliability and validity, and statistical validity will be relatively brief and is generally confined to this chapter. However, we also discuss these types of validity in more detail when they 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. Since 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 "better" 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 to teachers with less education, teachers' additional education "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. We discuss in subsequent chapters 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 words 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.

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 all Temporary Assistance for Needy Families (TANF) programs in the United States, would probably question the external validity of a study of TANF programs in Nebraska (no matter whether it was 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.

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. We discuss each of these in more detail below. It is important to note that measurement reliability and validity refer not just to outcome or output variables, but also to program variables, as well as other variables that the evaluation measures.

Finally, statistical validity refers to the accuracy with which random effects are separated from systematic effects. Measurement reliability and validity respectively pertain to the relative absence of random and systematic error in single variables. 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 pertains to relations between two (or more) variables; measurement issues usually pertain to variables considered one at a time.

## Types of Validity: Threats and Simple Remedies

With one exception, we discuss each type of validity in greater detail below. We reserve the entire next chapter to discussion of threats to internal validity, and the three following chapters 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 may 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 is about 200,000 members, she might select them so 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 Mary-

land, and he might decide that he cannot study all of them. Instead he decides on a representative sample, while 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; it is *not* the total number of students in the 750 schools.

Similarly, suppose a researcher wants want 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 so 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 will 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, and 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, 30, are not likely to be representative. What makes random samples representative is their size: large random samples of 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, no matter whether the population from which they are being selected is 30, 300, 3,000, 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 Interior 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. While 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.

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 just said, small random samples will 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 10,000 units, pretend to throw that unit back into 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 really do not toss the units back. So the probability of the first unit is 1/10,000 = .0001. The probability of the second unit is 1/9,999, which is only very slightly greater than .0001. The probability of the third unit is also only very 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 random sample of states from the population of fifty states will not be random. If a researcher aims for, say, a sample of thirty, the probability of the first unit is 1/50 = .02; the probability of the second is 1/49 = .0204; the probability of 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 they are being studied, their awareness often distorts their behavior, so 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 if 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, so 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 1930s 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 if the workers 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."

It would seem that the remedy for the problem of studying sensitized units is straightforward: do not tell people that they are being studied. While 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 cients, 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. While there are some exceptions to this rule,[6] 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 so that it is not representative of what would be observed outside of 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 we 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, or send them a survey by mail or e-mail, or visit their homes for a face-to-face interview. While the researcher selects the intended sample, the respondents select themselves, the actual sample; in effect, they are volunteers. Typical response rates are 70 percent or less. Even the response rate to the national

census (which is not a sample; it is a tally of observations from the entire population) is only about 70 percent. 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, National Institutes of Health (NIH) offers summer internships to healthy college biology majors to go to Washington, 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 8 discusses in considerable detail the steps that researchers can take to increase response rates to surveys, and we 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 big schools, small schools, state schools, expensive private schools, U.S. schools, French schools, 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 research 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 not only means that external validity is not possible but also can reduce 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 see below). For instance, African-Americans are 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 make sure 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 (no matter whether the claim is "vouchers work" or "vouchers do not work") might be externally valid for the larger subgroup (those who are not African-American), but it will be less so 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 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 to 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. For example, readers know from personal experience that there is an element of randomness in our performance on tests; usually you do well, sometimes you do not. In sports, sometimes you hit the ball or basket, but usually you do not. Similarly, the reason that researchers cannot generalize from small samples is not only that the small sample is unlikely to be representative of the larger population to which the researcher seeks to generalize. Such generalizations are prone to random error. If a researcher interviews one or two or ten people, she should know that that is too small a sample to characterize the larger population, not only because it is unrepresentative but also because it is not likely to be random. These are all aspects of statistical validity.

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 such as these.

## Sources

There are three sources of randomness in observational studies, no matter whether they are descriptive or causal. There is randomness in sampling, in measurement, and in human behavior. Consider, first, sampling as a source of random error. Recall that we have already related random sampling to external validity. Specifically, we said 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.[7] We note here that the probability of accuracy increases as the sample size increases, but 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 that.

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, so 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 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 one that is really there (the signal). 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 solely interested in estimating population characteristics 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 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 (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, 1,000 observations might be better, but not necessarily 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 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 pro-

gram 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 where the number of observations can only be one, 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 has been operating for 10 years, then the potential *N* is 10 years x 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), to that of a different program serving homeless women in the same city, using monthly data for the same period. Now, the *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, 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, the *N* is 120. If the researcher can collect similar data on, say, 100 homeless women in a different program in the same city, the *N* now becomes 220.

The point is that what looked originally like a study with one observation can be extended over time or, by looking at entities within a single unit, examined at a micro level, or both, simultaneously increasing both its statistical and external validity. It may also be possible to increase the *N* by adding another set of observations on individuals served by a different, comparable shelter in the same city, providing a comparison for the focal shelter 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. They are 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 10 might be a fluke, the evaluation would be more confident 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 variable like educational achievement or "social adjustment." For example, because of randomness in school performance measures, a school in a study might score low on one day, but if the same test were given in the same school the next day, the school 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 the students in the school, or repeated measures of different tests at the same school, or 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 fluctuating around the true mean. Thus, in the case of a single variable, especially when it is abstract, the best way to reduce randomness in observations or scoring is to have repeated measures or observations.

As another example, consider the design of standard educational achievement tests, such as the Scholastic Aptitude Test (SAT) or Graduate Record Examinations (GREs). Why are these tests so long? Asking multiple questions about the same basic concept increases the reliability of the test (with diminishing returns). 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 quicker but much "noisier" about a student's true performance than a fifty-item final exam. In fact, randomness in individual-level measures (the example in this paragraph) is usually far greater than randomness in collective-level or aggregate data (such as the school, discussed in the previous paragraph), but it does not disappear, especially when the concept to be measured is abstract.

We discuss the problem of random measurement error in more detail below. However, the point here is that random error in measures reduces statistical validity; the two concepts are related, because randomness in measurement introduces "noise," a source of statistical error. As we point out in the discussion of random measurement error, 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, more indicators reduce random error in the measurement of abstract concepts. With diminishing returns, multiple 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, 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, 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.[8] We discuss the implications of random responses (in tests, surveys, and other measures) below, in our discussion of measurement reliability and in our discussion 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 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. They 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.

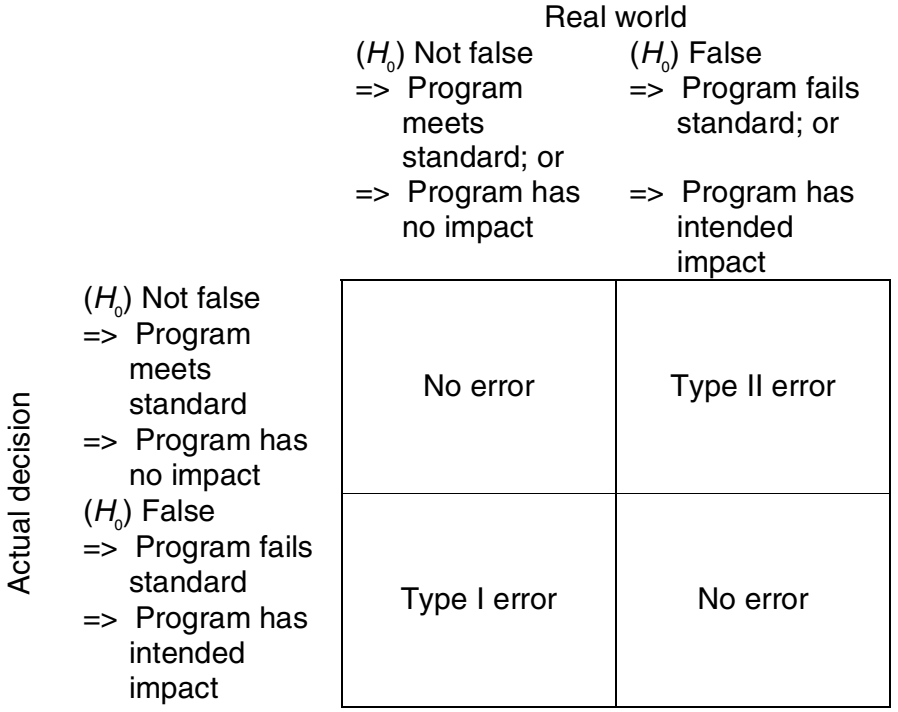
If the stochastic component is too large, it will be more difficult to separate any systematic impact of job training on wages from random patterns. We discuss below (in the chapters on research design) how researchers can make use of pretest scores to reduce the random component of human behavior without having to undertake the impossible burden of explaining it. The basic idea is that the best predictor of a person's wages at some time, *t,* is to know what that person's wages were at a previous time, $t - 1$. Researchers take advantage of the predictability (i.e., stability) of most people's behavior to reduce the stochastic component. Predictability does not really *explain* why some people earn higher wages than others do. But 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. We postpone that discussion to our 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. Type I error occurs when a null hypothesis is rejected when it is actually true; 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, we characterize each type of error; then we provide a simple illustration of how large samples can reduce the chance of each.[9]

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 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, has a 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 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.

Figure 3.1 **Statistical Errors in Hypothesis Tests**




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 in 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 3.1 depicts the evaluator's dilemma.

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 we said above, 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 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 one is "really" there.

No study is 100 percent valid, but some 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, once 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 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 if 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 to 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 Errors of Type I and II

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.[10] While that 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 to 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 to 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 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, on the other side, 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, these Type II errors will eventually encourage more (severe) crimes and 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 more costly.

In a murder case, the Type I error may arguably 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.[11] However, on the other side, if the jury rules that the defendant is guilty and puts him on death row, 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 puts 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.[12] Later, we will see that using "efficient" statistics and statistical tests can also reduce the probability of both Type I and II error. 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 accurate. In statistics, more accurate 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 below) 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 7 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 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 (vague) "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.[13]

However, a simple 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 one not to reject the null, the evaluator, in step two, 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 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.[14]

In a similar vein, Jeff Gill suggests paying attention to the confidence interval of a parameter estimate.[15] Confidence intervals decrease as the sample size increases, which is analogous to increasing the power of a null hypothesis test, where 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. We turn next to the critical issue of measurement.

## Measurement Reliability and Validity

### Introduction

Valid descriptions of program inputs, outputs and/or outcomes, and valid assessments of program impact requires 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 is more or less participatory and valid measures of output that represent productivity levels.

Abstract concepts such as 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 3.2 **The Basic Measurement Model**

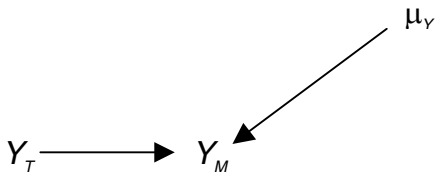



Figure 3.2 shows how $Y_T$ is related to $Y_M$. In this diagram, the measured scores ($Y_M$) are determined by a random component ($\mu_Y$) and a systematic, or nonrandom, component ($Y_T$). If most of $Y_M$ is comprised of $\mu_Y$, then $Y_M$ is a noisy measure of $Y_T$, with considerable RME. On the other hand, 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 in respect to NRME. Specifically, we write $Y_M$ as a linear function of both $Y_T$ and $\mu_Y$:

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

In this formulation, if the expected value of $\mu_Y$ 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, if $E(\alpha) = 0$ 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: $Y_M = 0 + 1^* Y_T + \mu_Y$. 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.[16]

Below, we discuss examples of both kinds of measurement error, 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 such as 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: $E(\mu_Y) \approx 0$. 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 that it registers a 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 or error in the intercept; and systematic 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. We have already presented some examples of measurement procedures that are likely to contain random components:

- 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 152. 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 152. 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 an SAT score is not as reliable an indicator of academic achievement because a student taking the SAT twice in a short period of time 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.

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.[17] 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 random measurement error, 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; we discuss nonrandom measurement errors below.) 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 (for example, those that result in death or hospitalization) are more likely to be reported. Thus data on airline crashes are probably more reliable than data on falling down stairs.

Random measurement error 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 record keeping, 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 effects. First, unreliable measures of outcome variables reduce statistical validity, thus raising the likelihood of both Type I and II errors. We discuss this further in Chapter 6 on nonexperimental design. 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.

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 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. We prove this statement in Chapter 7, where we discuss how to measure 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 tap 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, since 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, or 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.

## 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 *non*random measurement error (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 that 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 = $x$ perform more poorly on these tests than nonminorities, so that their observed score < $x$. 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 and test scores.

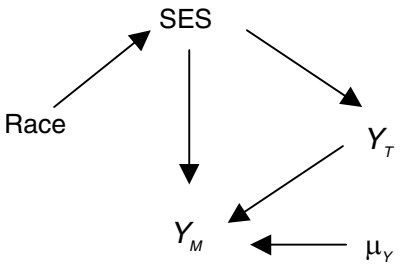
If the allegation were true, it would be an example of systematic error. Even though there is no direct connection between underlying true scores and race, racial minorities in the United States typically come from families with low financial and human capital assets. One consequence is low measured test scores, as a result of low assets, not race. Using the basic measurement model to represent this allegation, where $Y_M = \alpha + \beta Y_T + \mu_Y$, not only is there RME, but also E ($\beta$) < 1, unless race and capital assets (often measured by indicators of socioeconomic status, or SES) are accounted for. In this example, we assume $E(\alpha) = 0$; there is no consistent or correlated 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 test scores. Figure 3.3 represents this dilemma. Unadjusted GREs, SATs, PSATs, and other standardized test scores do not account for these alleged sources of systematic, or correlated, bias.[18]

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 (not all) survey items depending on whether the interviewer is the same race as the respondent.[19]

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,

Figure 3.3 **The Model of Systematic Nonrandom Measurement Error**




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 on-road 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 not only examples of face invalidity; they are also examples of consistent 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.

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 on-road 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? These questions raise issues of predictive validity. They are also instances of *consistent or 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, no matter 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. We will discuss these issues further when we discuss threats to internal validity in the next chapter and in Chapter 7. 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 2 on performance measurement also stressed the importance of multiple indicators. Chapter 8 on surveys briefly introduces factor analysis as a tool for assessing the validity of measurement procedures, 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 7, 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 topic of this chapter, and 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 GPRA 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: can you 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 error)
  - Are the measures likely to be correlated with factors that do not reflect the underlying concept? What factors? (correlated error)

# 4 Internal Validity

## The Logic of Internal Validity

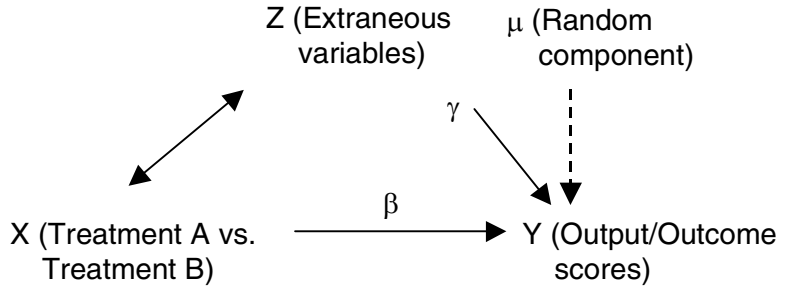
Internal validity is critical to the defensibility of impact evaluations. External validity, statistical validity, and measurement reliability and validity pertain to all program evaluations, no matter whether they are primarily descriptive or causal. In contrast, internal validity pertains only to impact or causal evaluations, and it is key to their credibility, because it is defined as the accuracy of the causal claim. If an evaluator claims that program *X* (e.g., welfare reform) causes outcome *Y* (e.g., reduced welfare caseloads), and if that claim is accurate, then it is said to be an internally valid claim. Similarly, if an evaluator claims that program *X* (e.g., legal political action committee (PAC) contributions to members of the U.S. Congress) does *not* cause outcome *Y* (e.g., how members vote on bills relevant to the PAC), and if that claim of no impact is accurate, then the claim is said to be internally valid.[1]

Logically, the problem of assessing internal validity can be broken into the components of a claim like this one: program *X* caused output or outcome *Y* to change by a certain amount (which could be positive, zero, or negative impact). There are three sources of any observed change (even if it is zero) in *Y:*

Observed level or change in level of *Y*
= Effect of intervention or treatment type or level (*X*)
+ Effects of other systematic processes (confounded or extraneous factors and related design effects of threats to external and measurement reliability or validity) (*Z*)
+ Effects of stochastic or random processes (threats to statistical validity and related design effects of threats to external and measurement reliability) ($\mu$)[2]

We can write this relationship as an equation: $Y = \alpha + \beta X + \gamma Z + \mu$. In the equation, the constant $\alpha$ is the value of *Y* if the program level is zero and other systematic processes are also at a level of zero. Theoretically, it stands for the base value of *Y* that would be observed if there was no treatment. However, only under limited circumstances could it be regarded as a counterfactual. The variable *X* is the level or dosage or amount of the program or intervention (e.g., before or after welfare reform, amount of PAC contribution); $\beta$ represents the actual magnitude of difference that a one-unit change in *X* makes on *Y,* net of *Z.* It is what we seek to find out. The next term, *Z,* stands for a summary of the level or amount of extraneous or confounding, but known and measured, factors; $\gamma$ represents the actual magnitude of effect that these factors have on *Y,* net of the intervention. Finally, $\mu$ is a summary of all the "stuff" that we left out because we do not know about it or because we cannot measure it.

Figure 4.1 **Sources of Variation in *Y*: The Picture**




The $\mu$ term has no direct measure; we hope that it is a random term that summarizes these omitted, unmeasured factors. Our central interest is in internally valid estimates of β.

We can also summarize the equation as a graphic, shown in Figure 4.1. The graphic clarifies the possibility that *X* and *Z,* the treatment (i.e., the intervention), and the confounding (or extraneous) variables, respectively, can be related to one another, in no particular direction, and therefore difficult to disentangle. The central problem of internally valid program evaluation is to do just that: to estimate β net of, or independent of, *Z,* which represents all variables that affect *Y* and are also related to *X*.

As an example, consider the case of welfare reform. Let *Y* be the measure of an important outcome: the number of months that the target families are on welfare. Let *X* be a measure of the intervention: whether the target families were receiving the traditional welfare program with no work requirements ($X = 0$) or a newer form, which requires recipients to work ($X = 1$). The coefficient β would be the difference that the two kinds of welfare program make on *Y;* β is unknown but it is potentially knowable, and the purpose of our evaluation is to estimate its value as validly as possible. Theoretically, that value could be 0, +, or –. A value of zero would mean that being on welfare without ($X = 0$) or with ($X = 1$) a work requirement makes no difference in the number of months on welfare (*Y*). A positive sign would mean that the new program increases the amount of time that the family remains on welfare. A negative sign would mean that the new program as a condition for receiving the benefit reduces the amount of time that the average family remains on welfare. *Z* is a summary measure that includes all the other factors that are related to being on some kind of welfare and that also affect how much time the family is likely to be on welfare. These factors include the family members' education, race, health, age, place of residence, and so on; the coefficient γ is a summary of their impact. Finally, $\mu$ represents unmeasured (and unmeasurable) factors that affect *Y* that are not included in *X* and *Z,* which are measured in our evaluation study.

As another example, consider estimating the impact of PAC contributions on voting behavior in Congress. We want to evaluate whether the receipt of money causes members of Congress to alter their vote, changing it to reflect the preferences of the PAC donor rather than those of their constituents. Suppose that the bills at issue are ones the PAC favors. We suspect that β is nonzero and positive. *Y* is a measure of how members in Congress voted on the target bills; *X* is a measure of the amount of PAC contributions from the suspect groups. To estimate the net impact of *X* on *Y* (that is, to accurately estimate β), we need to account for all the other factors (*Z*) that affect how members vote that are also related to the receipt of PAC money (*X*). These factors might include

party, ideology, constituency characteristics, and so on. To the extent that we fail to account for these factors, our estimate of β will not be accurate, because it will not separate the impact of PAC money (*X*) from that of the related, confounding variables (summarized by *Z*). That is, it will be internally invalid. We will also need to worry about the last term, $\mu$; it is a summary of everything else, besides *X* and *Z,* that affects *Y.* We do not measure the components of $\mu$; we do hope that their overall impact averages out to be zero. We know that we cannot eliminate this random component from our design, but, to improve the statistical validity of our estimate of β, we want to minimize the random component of our evaluation design.

These two examples of causal claims come from the United States. But program evaluation is critical to the study of international development, and evaluations of program impact are at the center of this growing subfield. Many interventions in developing nations focus on health, education, and financial services. For example, one study examined the impact of health intervention programs in Bangladesh (*X*) on child mortality (*Y*), controlling for, among other variables, mother's education (*Z*). (In this study, mother's education is not only a confounding variable; it also interacts with *X,* since the impact of *X* on *Y* depends on the level of *Z.* The health intervention reduced mortality the most when the mother's education was low.)[3]

Another study, also in Bangladesh, examined the impact of participation (*X*) in microcredit programs (such as those sponsored by the Grameen Bank) on labor supply, schooling, household expenditure, and assets ($Y_1 \ldots Y_4$), controlling for many (*k*) individual and village-level characteristics ($Z_1 \ldots Z_K$). Two of the control variables in this study (gender and income) also interact with the program (*X*), since the amount of impact of the program depends on income and gender. The program appears to be more effective (the estimate of β is greater) for poor women than for other groups.[4]

There are many evaluations of education interventions. For example, one examines the impact of improvements in school quality (*X*) on school attendance (*Y*) in Kenya, while another estimates the impact of an expansion in school supply (*X*) on youth wages (*Y*) in Indonesia.[5] A third studies the impact of extending tuition-free education (*X*) on various labor market outcomes for youth (*Y*) in Taiwan.[6] All of these studies control for numerous confounding variables (*Z*).

In each of these examples, and in causal program evaluations in general, what we really care about is the net effect of *X* on *Y.* To estimate that net effect, we have to rule out the other systematic factors, besides *X,* which affect *Y,* as well as the random ones ($\mu$). Failing to account for the components of *Z* that affect *Y* and that also are related to the intervention (*X*) threatens the internal validity of the estimate of $\beta$. We also need to rule out the possibility that we may be erroneously attributing a systematic effect (that is, the claim that *X,* the program or policy, the alleged causal agent, is systematically related to *Y,* the output or outcome) to a random one ($\mu$). We also need to rule out the possibility that a claim of *no* systematic effect (that is, a claim that *X* has only a random effect on *Y*) is really a systematic effect. Chapter 3 (briefly) discussed some of these design effects. That is, we saw in Chapter 3 that a causal claim of positive, negative, or zero impact may be inaccurate because it is not externally valid, because it is not statistically valid, or because it lacks measurement reliability and validity. Chapter 3 also pointed out that many aspects of threats to external validity and measurement reliability are important problems because they are threats to statistical validity. In the equation $Y = \alpha + \beta X + \gamma Z + \mu$, statistical validity is partly captured by the $\mu$ term, and we have already discussed many of these threats in general. (We discuss others in Chapter 7 on nonexperimental design.)

We point out at the end of this chapter that many other aspects of external invalidity and measurement unreliability and invalidity can best be regarded as threats to internal validity. Fur-

ther, all the threats to internal validity can be regarded as *Z*-variables that, if they are ignored, may be confounded with the program variables under study (the *X*-variables) in that they also have a systematic effect on the output or outcome variables (*Y*). Failure to consider these threats (*Z*) by separating them from *X* may make the program (*X*) look effective when it is not or may make the program look ineffective when it is actually effective.

It is important to reiterate that no study is 100 percent valid. No study can simultaneously rule out every threat to each of the four kinds of validity, and no study can ever entirely rule out even one of the four threats to validity. However, some studies are clearly more valid than others, in that they do a better job of minimizing threats to validity. We focus here on numerous threats to internal validity. By providing a checklist, evaluators can anticipate these threats, designing their study to either account for or fend off these threats as well as possible. The poorest studies do not consider these threats at the design stage or analysis phase at all.

## Making Comparisons: Cross Sections and Time Series

Before considering these threats to internal validity, recall from Chapter 1 that all causal studies use contemporaneous and/or retrospective observations. That is, they are based on observations of ongoing and/or previous activities. All causal analysis also requires comparison. Without comparison, there can be no counterfactual: what would have happened to the outcome (*Y*) if there were no intervention (*X*) or if the intervention had been different?

There are two basic ways to make empirical comparisons. One type of comparison is a cross-section (CS) study; the other type is a time-series (TS) study. Consistent with the idea that no causal claim can be made if there is no comparison, both basic types of design entail comparison.

The *cross-section design* compares two or more similar units, one with the program and one without the program (or one with a different program), at the same time. Thus, one might compare standardized test scores in 1998 in a school district that has school choice with scores in 1998 in a similar district that does not allow school choice. Alternatively, one might examine test scores in two comparable school districts at the same time, one district having extensive school choice while the other has minimal choice. Or one might examine three districts, one providing no choice, the second providing a little, and the third providing an extensive school choice program. These would all be CS designs.

By contrast, time-series designs compare the same unit at two (or more) points in time; the earlier observations represent preprogram outcomes, while the later observations represent observations during the program. Or the comparison could represent observations before and after a program change. For example, one might compare the number of crimes reported by citizens to police before a community-policing program went into effect to the number reported one year after. This is a simple before-and-after comparison. Another type of TS design might entail more than two data points. For instance, one might compare trends over a period of time in measures of public school outputs or outcomes (e.g., teacher-to-student ratios, standardized achievement scores), before a property tax limit began to the trend in the same measures for a period of time after. Each of these is a different kind of TS design. In addition to examining outcomes before and after the inception of a program, it is also possible to examine trends (or levels) before and after a change in the amount of program resources. For example, one could examine crime levels in a city before and after an increase in the number of police on the street.

Some designs combine cross sections and time series. For the purposes of introducing the main threats to internal validity in both kinds of designs, we will not discuss mixed designs here, but we do consider combination designs in Chapters 5 and 6. All designs try to separate the impact of an intervention ($X$) from extraneous or confounding systematic factors ($Z$), and from random influences ($\mu$). The preceding chapter considered many sources of random influences. This chapter considers the problem of systematic factors ($Z$) related to $X$ that also affect $Y$. These factors represent threats to *internal* validity. It is not possible to say that one design (e.g., CS) is better than the other (e.g., TS) in reducing threats to internal validity. Rather, both have advantages and disadvantages. Specifically, they are each subject to different threats to their internal validity.

## Threats to Internal Validity

### *History or Intervening Events*

The threat of "history" or "intervening events" applies to all time-series studies. That is, many events (included in $Z$) besides changes in the program that is being studied ($X$) could have happened during the period of the study. It is always possible that the observed change in $Y$ may be due to the intervening event ($Z$) and not to change in the program being studied ($X$). For example, if a researcher examines the number of crimes reported to police ($Y$) by citizens both one year before the implementation of a community policing program ($X$) and one year after, the observed change in reporting behavior (or the lack thereof) may not be attributable solely to this program. The reason might be that half a year ago, the government fixed all the streetlights ($Z_1$), so the level of crime went down and there was just less crime to report. Or maybe a new business opened up in town ($Z_2$), expanding employment opportunities and reducing the incentive for criminal activity. Similarly, if a researcher is studying the impact on graduation rates of a new work-study program for at-risk high school students by comparing graduation rates before the program to those after, what the researcher observes could be attributable to the new program. But it could also be explained by other events, such as a new principal or a change in the curriculum that affected the school during the same period. The point is that an extraneous event, and not the one under study—the new work-study program—could account for the observed change (even if it is no change) in graduation rates and other output or outcome variables.

These intervening events of history may be difficult if not impossible to disentangle from the treatment that is being studied. We suggest some strategies for doing so in our discussion of specific research designs in Chapters 5, 6, and 7. But the greatest threat is the evaluator's failure even to consider history or intervening events (a $Z$ variable) as a possible source of systematic change that is extraneous to the program under study ($X$) and that may be confounded with the program if it is totally ignored.

### *Maturation or Secular Change or Trends*

In time-series designs, besides the intervention due to the program under study ($X$), outputs or outcomes ($Y$) may be systematically affected by long-term, underlying trends ($Z$), which can also be characterized as maturation or secular change. We may observe a change in $Y$ before and after a program change, but the change may be due to a long-term underlying trend, not to the program. For example, decreasing crime rates may not be due to a certain anticrime program, but occur

instead because people are aging. Older people are less likely to commit crimes, so crime rates would go down anyway, even without the anticrime program. Similarly, changes in standardized achievement tests (*Y*) in schools affected by a new education initiative (*X*) may also be affected by long-term trends in the demographic composition of the school district (*Z*). If the school district has a growing population of at-risk minorities or immigrants, school test scores (*Y*) may be dropping, and it is important to separate this underlying downward trend (*Z*) from the impact, if any, of the education initiative (*X*). Another example illustrates why this threat is also referred to as "maturation." Suppose that an evaluator is studying the impact of a new curriculum on third-graders. He examines reading scores in September, before the introduction of the new curriculum, and again in June, after the new curriculum. The scores (*Y*) improve. Maybe the new curriculum (*X*) accounts for the improvement, but maybe the cause is maturation (Z). Maturation in this case means that young students age, even between fall and spring during an academic year, and read faster and with more comprehension because they have additional experience and greater developmental readiness. Consequently, some (or all) of the improvement might have happened anyway, regardless of the curriculum change. Thus, internally valid designs must always try to disentangle the program (*X*) from confounding factors (*Z*), such as maturation or underlying trends in these examples.

Some time-series designs make it easier to reduce threats due to long-term trends or maturation than others. For example, it is easier to reduce these threats in designs that have observations at many different points in time. It is impossible to deal with this threat in simple before-and-after designs with only one "before" observation and one "after" observation. In that case, the program (*X*) is measured as "before" and "after," and the trend (*Z*) is measured as the very same two data points. But if there are many pre- and postprogram observations, it is easier to reduce this threat by separating the underlying trend from the inception of the program. We discuss this issue in more detail in subsequent chapters. We also point out in the chapter on nonexperiments that the threat of maturation or trends corresponds to the problem of auto- or serially correlated data and that it is a threat to both internal validity and to statistical validity. But the greatest threat is for evaluators who use time-series designs to ignore this threat entirely.

## Testing

"Testing" refers to instances in which the method of measuring the outcome *Y* can affect what is observed. This is common in the social sciences. Observing a rock will not change its behavior; observing me will change my behavior, if I see you looking at me. These kinds of measurement techniques are called obtrusive measures. Not all measures are obtrusive. For example, collecting administrative data about my age, years of employment, place of work, rank, and salary will not affect my behavior, especially if I do not know about it (more about this issue later). Surveys and tests, however, can be obtrusive. Responding to a survey or taking a test can change my behavior: surveys may make me aware of new issues, and I learn by taking a test. Testing is a threat to internal validity in both CS and TS designs.

Obtrusive measures are a particular problem in TS evaluations. In TS studies, when repeated measures are obtrusive, taking the test (*Z,* or the pretest score on the output or outcome measure, $Y_{t-1}$), and not the intervention or program treatment (*X*), may cause the outcome (*Y,* or the posttest measure $Y_t$) to change.[7] We will be unable to tell whether the observed change in *Y* is caused by the obtrusive pretest measurement (*Z*) rather than (or in addition to) the intervention or treatment (*X*). We saw in Chapter 3 that taking tests is a threat to external validity. We see now that obtrusive measurement is also a threat to internal validity in time series, and it can be difficult to correct.

An example will clarify. Suppose that a manager, concerned about productivity, surveys the employees in his organization to assess their morale, their networking, and their commitment to the organization's mission. Finding from the survey that morale, networking, and commitment are not at the level he thinks they should be, he reorganizes the office to make it less hierarchical. Several months later, he resurveys the employees and is glad to see that morale, networking, and commitment have improved. Can the manager attribute the observed change in $Y$ to the reorganization ($X$)? Not necessarily. Some (or all) of the change could have been due to the very act of pretesting to get the baseline data on $Y$. Just as in the famed Hawthorne study, pretesting may have sent a signal to the employees that the manager "cared."[8] The signal from the pretest (called either $Y_{t-1}$ or $Z$), and not the actual reorganization ($X$), may have accounted for the improved scores on the output measure ($Y_t$).

Testing effects are difficult to avoid in a TS design when measures of outcome or output are obtrusive. One way to minimize these effects is to use multiple indicators. For example, in addition to the survey, the manager could use other indicators of office morale and commitment, such as objective indicators from administrative data on productivity and absenteeism, or observations of how many people arrive early and work late. These measures are less obtrusive and could supplement the survey. While the survey may be a more direct measure of what the manager wants to know, its disadvantage is that it is obtrusive and therefore subject to testing effects.

Testing is also a threat to internal validity in CS designs. The classic example of this threat is the use of a placebo in cross-section studies of pharmaceuticals. For example, if one were to compare Newpill to "no treatment" ($X$), the process of administering Newpill ($Z$) may be inseparable from ingesting the pill itself ($X$). Administering the prescription requires physician intervention before the pill is ingested. In looking at the impact of $X$ on $Y$ (the duration or severity of a disease), it would be impossible to disentangle $X$ (the pill) and $Z$ (physician intervention). Consequently, medical researchers resort to the use of a placebo. One group of patients gets Newpill; the other gets Placebo or Fakepill. Both groups get "treated" with physician intervention. In effect, the use of a placebo turns a testing threat to internal validity into a testing threat to external validity, which is usually considered less severe.

## Instrumentation

In TS or CS, change in the calibration of the measurement procedure or instrument ($Z$) may partly or entirely cause the outcome ($Y$) to change, rather than the treatment ($X$). For example, in a TS study, if one observes a decrease in program costs ($Y_t$) after the implementation of a new technology ($X$), the observed decrease ($Y_t - Y_{t-1}$) may be due to the new way that costs are calculated ($Z$), instead of to program ($X$) effects. In other words, if $Y_{t-1}$ is measured differently than $Y_t$, then instrumentation becomes a threat to the validity of a causal claim about the impact of the new technology, because it is impossible to disentangle the new measurement procedure from the new technology.

Similarly, in CS studies, suppose that an evaluator (or a politician) claims that waste collection costs in one community that has privatized the service are less than those in a similar, nearby community, where the government runs the service. If the costs are measured differently in the two communities, some or all of the observed difference in costs ($Y$) may be attributable to differences in how costs are measured ($Z$), and not to whether the program is administered publicly or privately ($X$).

Minimizing this threat clearly requires careful attention to how outcome or output variables are measured over time and between or among the units of analysis at a single point in time.

## Regression Artifacts or Regression to the Mean

Subjects are sometimes selected for treatment because of extreme pretest scores on an output or outcome measure ($Y_{t-1}$). In this case, using a time series study, an observed change or difference in outcome ($Y_t$) may be observed partly or entirely because of the random tendency of extreme scores to return to their normal values. For example, suppose you took a Scholastic Aptitude Test (SAT) test and got an unexpectedly awful score ($Y_{t-1}$ = low). What would you do? Such an extremely poor score (one that is way below your usual or expected score) would increase the probability that you would sign up (and pay) for an SAT-preparation program. After completing the program, your score improves ($Y_t > Y_{t-1}$ = low). Is the improvement attributable to the SAT-prep program ($X$) or to a regression artifact ($Z$, or, equivalently, $Y_{t-1}$ = low)? Regression artifacts refer to the tendency of extreme scores to return, by chance, to their mean. The more extreme the score, the more likely it is the next time to bounce back to its mean. In fact, the higher the score, relative to its usual mean, the more likely it is to fall. Similarly, the lower the score, relative to its usual mean, the more likely it is to improve the next time. So the next time, after the preparation program, it is very likely that your score would improve. Is the improvement really due to the effect of the program? Maybe, but it might be simply a return of a randomly extreme score to its normal level. We note that those who score unexpectedly high on the SAT are unlikely to sign up for an SAT-preparation program.

Another example concerns sports teams. Sports teams are especially prone to change their manager after a really poor season, but any observed improvement in performance after that may be attributable to the chance return to normal that is expected after a worse than normal season, and not to the efforts of the new manager. As another example, municipal police forces tend to target police to areas that have spikes in crime rates. There is an element of randomness in crime rates (due to measurement error or randomness in human behavior). After a period, once the extra police patrols have been assigned, the crime rate appears to drop. The police chief claims credit, but the effect could be partially or entirely due to a regression artifact, representing the tendency of extreme scores to return to their usual, mean value.

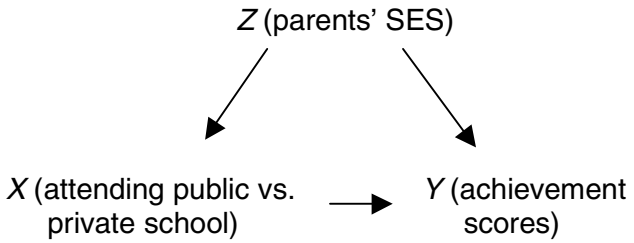
As a final set of examples, bosses (or teachers) often blow up in anger at employees (or students) who perform at suddenly poor levels, relative to their normal level. Soon after, the performance level improves. The boss (or teacher) concludes that anger works, but the improvement may simply represent the tendency of extreme scores to return to normal values. The flip side of this example is that bosses (or teachers) sometimes lavish praise on employees (or students) whose performance is suddenly exemplary. Soon after, the employee's (or student's) performance appears to decrease, returning to its normal level. The boss (or teacher) concludes that praise does not work, but this too may be an invalid inference. The apparent drop in performance may merely represent the tendency of randomly extreme scores to return to their normal level.

These are all examples of the problem of regression artifacts in TS studies.[9] As we see below, the best way to reduce the threat is to collect observations over many points in time, so that it is possible to separate normal from extreme scores. It is also possible to make use of randomness in scores by constructing a type of natural experiment, called the regression discontinuity design. We discuss these issues further in Chapters 5 and 6.

## Selection (Uncontrolled Selection)

In CS studies, when the groups to be compared differ on factors besides the treatment ($X$), then these differences ($Z_1 \ldots Z_k$) may account partly or entirely for the observed differences in out-

Figure 4.2 **Causal Model of Selection Threat**



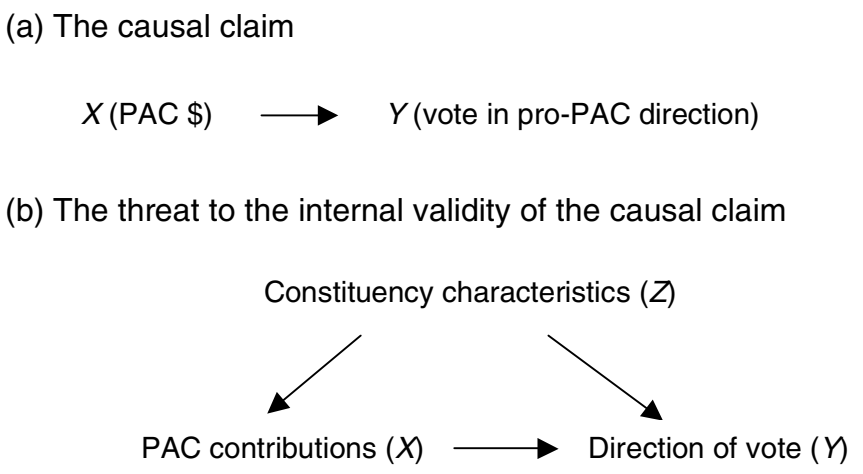

come (*Y*) between those who receive the treatment or program and those who do not. For example, private school students' performance on standard test scores is much better than that of students from public schools. Can we attribute the observed difference simply to attending different types of schools: private or public? Maybe. But students who attend private schools tend to be wealthier and to have parents who are more educated than those who attend public schools. Thus, some or all of the observed difference in achievement scores (*Y*) may be due to the tendency of wealthy, educated parents (*Z*) to select private rather than public schools (*X*) for their children. Thus, if one does not account for the selection effect, the impact of socioeconomic status (SES) (the *Z* variable in this case) on *Y,* the outcome, would be confounded with the type of school, which is the treatment (*X*) variable. Figure 4.2 is a graphic of the causal model that depicts this dilemma.[10]

Failure to account for variables (*Z*) that are related to the program or treatment (*X*) and that also affect the outcome or treatment (*Y*) will cause the evaluator to confound or mix up the impact of the program with that of the uncontrolled or unaccounted-for variable. The consequence is a conclusion about causal impact that is likely to be erroneous. In the example, ignoring SES differences between public and private school children is likely to result in overestimating the impact of private relative to public school education.

This is also a problem in comparing public schools to one another. For instance, if an evaluator simply compares one public school (say, PS 105) to another (say, PS 4) and finds that the achievement scores at PS 105 are higher than those at PS 4, she cannot simply conclude that PS 105 is a "better" school than PS 4. PS 105 may simply serve a higher SES group of students than PS 4. Some (or all) of the observed difference in outcome (*Y*) may be due to the difference in SES between the schools (*Z*), and not to any particular programmatic or management differences between them (*X*).

As a final example, many claim that Political Action Committee (or PAC) contributions (*X*) make legislators vote (*Y*) the way the contributors want. Symbolically, the claim is that *X* causes *Y,* which we diagram in the top panel of Figure 4.3. However, PAC contributions are not randomly distributed. Trade union PACs direct their contributions to representatives from districts with union members and liberal Democratic voters. Similarly, business PACs direct their contributions to representatives from districts with large corporations and Republican voters. Representatives from these districts are likely to vote the way the PAC wants them to not because of the PAC contribution, but because incumbents want to get reelected and therefore vote the way their constituents would want. The lower panel of Figure 4.3 represents the causal diagram for this scenario. It shows that some or all of the observed correlation between PAC contributions (*X*) and representatives' votes (*Y*) in the direction preferred by the PAC may be attributable to a third variable, constituency preferences (*Z*), and not to the PAC contributions (*X*).

Figure 4.3 **Selection Threat: Does Money Buy Votes?**

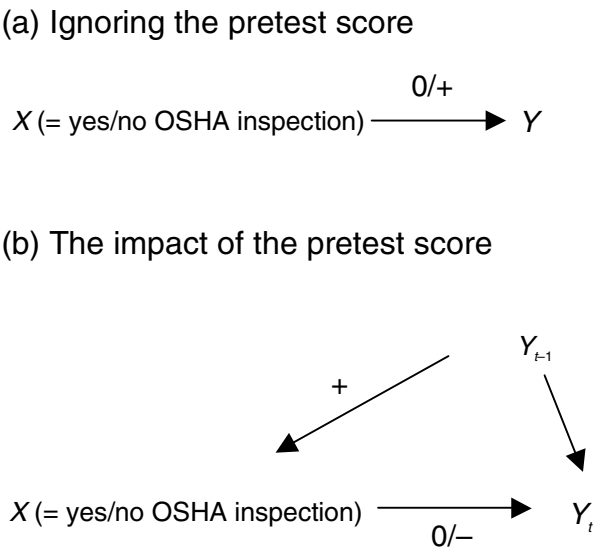



Ignoring the selection problem is equivalent to confusing causation and correlation. One cannot attribute correlation to causation except under very special circumstances. To do so is to invite internally invalid causal conclusions.

Selection problems are probably *the* dominant threat to internal validity in CS studies. There are thousands of examples of them. In fact, except when assignment is random, they are impossible to avoid in CS studies. Even if one controls for one selection threat (factor $Z_1$), there is always the possibility that another one lurks (factor $Z_2$). Even if one deals with ten selection threats, there is always the possibility that an eleventh or twelfth threat still looms. In fact, the presence of selection threats in most CS studies makes TS studies particularly desirable from the perspective of minimizing selection threats to internal validity. Many evaluators bemoan the absence of time-series data. Comparing one unit at time 1 to itself at time 2 is a useful way to hold constant all of these confounding ($Z$) factors that plague CS studies.[11] Often, however, TS data are completely unavailable or too sparse for statistical validity (that is, too few TS observations are available). So it is important to be aware of selection threats to internal validity, and to design CS studies to minimize these threats.

Uncontrolled (and unknown) pretest scores on outcome measures are also an important selection threat to internal validity in CS studies. This threat is analogous to the regression artifact in TS studies. For example, suppose that an evaluator compares workplace accident rates ($Y$) in firms that Occupational Safety and Health Administration (OSHA) inspects ($X$) to accident rates in firms that OSHA does not inspect. The researcher might find that the accident rate in inspected firms is higher than or even no different from the rate in firms that OSHA did not inspect. The diagram in the top panel of Figure 4.4 illustrates this apparent conclusion. The researcher then might infer that OSHA inspections fail to reduce accident rates (0 association) or even increase them (+ association). But suppose that, unknown to the researcher, OSHA targets inspections at firms with unusually high accident rates, as the lower panel of Figure 4.4 illustrates. Accident rates have a large random component, so some extremely high rates are likely to drop, by chance, to their usual expected rate. Thus, in the absence of information about how OSHA selects firms to inspect, and in the absence of information about the problem of regression arti-

Figure 4.4 **Selection Threats: The Case of Selection Based on Pretest Scores**




facts, the researcher might well conclude, erroneously, that inspections do not work. But this conclusion could be partly or entirely wrong, because it could be attributable to a regression effect rather than the systematic effect of the inspection program.

More generally, when selection is targeted at needy cases, we can regard previous values of output or outcome variables ($Y_{t-1}$) as a selection variable ($Z$) that needs to be accounted for in order to reduce an important threat to internal validity in CS studies. Using CS designs in these cases is, in many circumstances, a poor choice, because, by definition, simple CS studies have no pretest data, so that it is impossible to measure and adjust for the selection threat.[12]

## *Experimental Mortality or Attrition*

In TS studies, observed before-and-after differences in outcome scores ($Y$) may be partly or entirely attributable to a different group of respondents rather than to the treatment ($X$). For example, suppose the Department of Housing and Urban Development wanted to study the impact of housing vouchers on the recipients' housing location decisions. Suppose the researchers use a survey to compare where people live before they get the voucher to where they live afterward. The comparison would be valid only if the prevoucher respondents are the same as the postvoucher respondents. It is hard to follow up survey respondents, and it is particularly difficult when those respondents are poor and subject to the unpredictable forces of a shaky labor market. Thus those who respond at time 1 but not at time 2 could be a less well-off group than the more stable and more employable respondents who actually respond at both time 1 and time 2. The researchers might conclude that the vouchers improved housing outcomes, but maybe it is the difference between the SES ($Z$) of the respondents at time 2 and time 1, rather than the impact of the vouchers ($X$), that accounts for the change in $Y$. When respondents at time 1 drop out at time 2, they are called "attritters"; for the researcher, they represent a "death" (i.e., an experimental mortality) if they cannot be found.

Tracking changes over time in achievement test scores presents another example of an attrition threat. It may illustrate experimental addition, rather than attrition, but it is the same underlying phenomenon. If the test takers are changing, then the changing scores may measure not changes in performance, but changes in who is taking the test. There are many examples of this threat in TS studies of student achievement test scores. For example, one reason that SAT scores go down is that more students take the test, and the newest test takers tend to score more poorly than the previous ones. Thus, drops in test scores may not reflect poorly on the schools, but may entirely or partly reflect changes in the types of students who take the SATs. Similarly, school achievement test scores (*Y*) may change from one year to another not because of any response by the school (*X*) but because the demographics of the students in the school (*Z*) have changed.

Attrition itself is *not* a threat to internal validity. (It may be a threat to statistical or external validity.) However, as the examples above illustrate, if the attrition rate is related to the treatment variable (*X*) or to a confounding variable (*Z*) that is related to both *X* and *Y,* then attrition is a threat to internal validity.[13]

## Multiple Treatment Interference

In TS or CS studies, when one treatment (*X,* the treatment of interest) is confounded with another (*Z,* another treatment, but not the focal one), then it is impossible to separate the impact of one treatment from the other. For example, suppose that an evaluator, using a CS design, compares the impact of rigid versus flexible curricula (*X*) in the first grade in otherwise similar schools (or classrooms) on standardized achievement scores (*Y*). However, it turns out that all the schools (or classrooms) with flexible curricula also have very experienced teachers (*Z*), while the rigid curricula are taught by teachers with less experience. Then *X* (the treatment of interest) cannot be separated from another "treatment" (*Z*) that is not the central interest. Later, in Chapter 7, we see that this is an example of what is called multicollinearity. For now, it is an example of multiple treatment interference, and it is regarded as a threat to internal validity because the two treatments cannot be separated.

Suppose that an international aid agency used a TS design to evaluate the impact of its decision to decentralize the management of agricultural projects in a particular country in Africa. The agency might look at a measure of farm productivity (*Y*) for some time before projects were decentralized (*X*) and for a period afterward. However, at the same time projects were decentralized, project budgets were also cut (*Z*). It is thus impossible to separate the impact of *X* from that of *Z.* The two treatments go together, exemplifying the threat of multiple treatment interference in a TS design.

## Contamination

In CS studies, sometimes the separation between treatment groups and control groups is less than it should be. For example, if the control group receives some of the treatment, then the two groups are no longer distinct, and the treatment is said to be contaminated with the control. Similarly, if some elements in the treatment group do not receive the treatment, then the two groups are once again no longer distinct, and the treatment and control groups are still contaminated. Unrecognized, this overlap (i.e., contamination) may partly or entirely account for the observed between-treatment difference in outcomes (*Y*). For example, in a study of drug prevention programs in

schools, suppose that one compares the attitudes toward risky behavior (*Y*) of students in the fourth-grade class that received the program (*X*) to similar fourth-grade students in the same school who did not experience the program in their class. Nothing prevents students who received the formal program from talking to those who did not receive the formal program. As a result, there is contamination in that some students who did not receive formal training are nonetheless exposed to elements of the program. Similarly, those in the formal treatment group may have chosen not to listen to the antidrug instructor or were absent; the treated group is effectively contaminated with students who were really not treated. Thus, a conclusion about the difference in outcome due to receiving versus not receiving the treatment may not be valid, if those who supposedly received no treatment actually received some treatment or if those who supposedly received treatment actually received none. Using our standard notation, the "pure" treatment-or-no-treatment variable is *X,* the self-selected treatment is *Z,* which is often unmeasured, and the outcome is *Y.* (This is also an example of nonrandom measurement error, as the actual measure of the treatment as a yes-no variable is clearly not valid.)

As another example of contamination as a selection threat, consider a study comparing the impact of classroom education to on-the-job training (OJT) (*X*) on employment outcomes (*Y*). The researcher usually assumes that subjects in the study receive either one treatment or the other. But many who are assigned to OJT ($X = 1$) may also elect to take education training (*Z*) elsewhere, and those in the education group ($X = 0$) may also elect to take OJT elsewhere (*Z*). Some in treatment groups may opt for no training at all. Once again, the two treatments are not as different as the researcher assumes, since one group has clearly been contaminated with the type of treatment offered to the other group. It is invalid to assume that the groups are really as separate as they appear to be.

Since the measure of "type of training" as two mutually exclusive categories (OJT vs. education in the example, or $X = 1$ or 0) is clearly not valid, it represents a threat of contamination (and nonrandom measurement error) to internal validity. In the example, the actual level of training (which we conceptualize as a confounding factor) is self-selected, while the intended level of treatment is what is being studied and measured as *X.* It follows that, if possible, one remedy for contamination threats is to separate the intent-to-treat variable (*X*) from the self-selected treatment variable (*Z*). The distinction between *X* and *Z* in this case may be of more academic than practical interest. In the real world, where treatment (and nontreatment) cannot be forced, the program is an opportunity to receive service, or "intent to treat," not actual receipt of the service. Similarly, "no treatment" is not the same as "no service"; rather, it means that the specific service under study is intended to be withheld, but people cannot be prohibited from selecting it on their own. This is a common problem in experimental designs, and we discuss it further in Chapter 5.

## Summary

To summarize, the threats to the internal validity of TS designs are somewhat different from the threats to the internal validity of CS designs. Threats to TS studies alone include history or intervening events; maturation, secular change, or trends; testing; regression artifacts; and experimental mortality or attrition. Threats to cross-section studies alone include selection effects or uncontrolled selection and contamination. Instrumentation and multiple treatment interference are clearly threats to the internal validity of both kinds of designs.

**Type of Threat by Design Type**

| Threat | Design type |
|---|---|
| History or intervening events | TS |
| Maturation, secular change, or trends | TS |
| Obtrusive testing | TS + CS |
| Regression artifacts or regression to mean | TS |
| Selection or uncontrolled selection | CS |
| Contamination | CS |
| Experimental mortality or attrition | CS + TS |
| Instrumentation | CS + TS |
| Multiple treatment interference | CS + TS |

Most studies have multiple threats to internal validity, but it is clear that some have more threats than others. For example, simple comparison of outcomes between two groups, one that has received the treatment while the other has not, is almost always subject to selection threats. In fact, unless one has a controlled experiment, which we discuss in the next chapter, any CS design is subject to selection threats. While some CS designs are more vulnerable to this threat to internal validity than others are, some researchers argue that the pervasiveness of selection threats in nonrandomized CS designs (which is most of them) makes them less desirable than TS designs. However, as we have seen, TS designs also have special problems.

It is probably the case that no single approach and no single study can ever be causally conclusive, since threats to internal validity are pervasive. But this is not a reason to give up. Rather, when multiple studies, each with different threats to internal validity, all suggest the same causal conclusions, we can act as if the causal claim were upheld. For example, while it is technically true that no study has *proved* that "smoking causes cancer," multiple studies, each with different threats to internal validity, all suggest that such a conclusion cannot be rejected. In this case of multiple, individually imperfect designs with a consistent finding, it is probably wise to act as if the causal conclusion is in fact empirically warranted. (Note that we still do not say that the conclusion has been "proved." Mathematicians do proofs. Empirical social scientists do not.)

## Three Basic Research Designs

The issue of internal validity is so critical to impact evaluations that three different ways of reducing threats to internal validity define the three basic types of research designs that evaluators use to estimate program impact. Undoubtedly, the most commonly occurring threat to internal validity is self-selection. Consequently, each of the basic design types reduces the threat of selection in a distinctively different way. (See Table 4.1.)

The first type of design is the randomized field experiment (RFE). In the RFE, the evaluator randomly assigns units of analysis to a treatment group ($X = 1$) or to a nontreated (or control) group ($X = 0$), or to groups with different levels or types of treatment, to reduce selection and selection-related threats to internal validity.

Second is the quasi experiment (QE). To reduce threats to internal validity in the QE, the evaluator deliberately selects treatment and other groups so they are comparable or similar in as many respects as possible with respect to confounding or extraneous factors ($Z$). The idea is to construct groups so that the only difference between the groups is in how much or whether they

Table 4.1

**Types of Research Designs, by Method of Reducing Threats to Internal Validity**

| Type of design | Method of reducing threats to internal validity |
|---|---|
| Randomized field experiment (RFE) | Random assignment by evaluator to program or no program, or to different program levels (e.g., high, medium, low). Evaluator compares outcomes between the groups. |
| Quasi experiment (QE) | |
| Cross-section (CS) | Evaluator selects groups with and without the program, or with different program levels. Groups are chosen so they are as similar as possible in all other respects, except the program. Evaluator compares outcomes between the comparable groups. |
| Time-series (TS) | Evaluator selects a target group, comparing outcomes before and after the implementation of the treatment. |
| Nonexperiment (NE) | Evaluator collects data on units of analysis that have experienced different levels of the program and compares outcomes in units with different levels of the programs, using statistical methods to control for other differences between the units. |

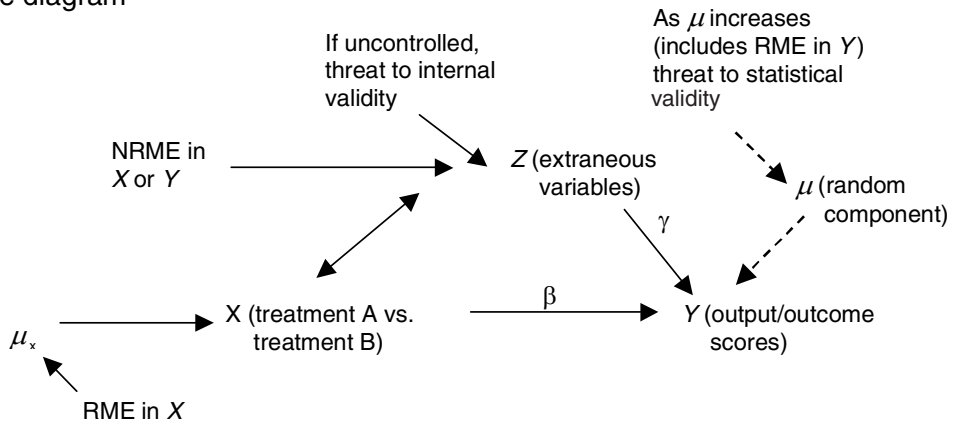
experience the program being evaluated. In one type of QE, the evaluator selects comparable groups at the same point in time, but one group has experienced the program (or experienced a high level of the program), while the other group has not. This is a cross-section quasi experiment (CS QE). In the CS QE, it is rare that the researcher determines whether the group experiences the program or what the program level will be. Usually, that determination has previously been made by political or administrative decision makers, and the evaluator examines the impact of the program *ex post* or retrospectively. Another type of QE examines one group over time, so that the before-treatment group is comparable to the after-treatment group, because the group is the same, save for the treatment. This is a time-series quasi experiment (TS QE). Some types of QEs use a mix of TS and CS comparisons.

The third type of design is the nonexperiment (NE). This design uses statistical controls to account for as many selection threats as possible. The main statistical tools for nonexperimental designs are multiple regression and related methods of multivariate statistical analysis. Finally, some studies use a mix of two or even all three types of designs. It is especially common for studies to supplement RFEs or QEs with nonexperimental (NE) design elements by adding statistical controls to reduce additional threats to internal validity that cannot be removed by selection or random assignment alone.

The remainder of this book discusses each of the three basic types of designs (RFE, QE, and NE). Even though it may be difficult to separate one type of design from the other, the three basic design types in Table 4.1 represent different approaches to establishing comparison, or to establishing a counterfactual. The goal is the same: to find out what would have happened to the outcome or output ($Y$) if the program ($X$) had not been implemented or had been implemented differently, net of the effects of the confounding influences of extraneous systematic factors (Z) and random effects ($\mu$). The choice of a design, however, depends on many factors, including the stage of the program (old or new), the type of intervention (targeted or general), and the resources

Figure 4.5 **The Causal Model Workhorse and Threats to Validity: Diagram and Statistical Model**

(a) The diagram




(b) The equation

$$Y = \alpha + \beta X + \gamma Z + \mu$$

for evaluation. The challenge facing the evaluator is to come up with the most rigorous or valid design (or mix of designs) that is feasible under the circumstances.

## Rethinking Validity: The Causal Model Workhorse

The previous chapter introduced the basic concepts of external validity, statistical validity, and measurement reliability and validity. This chapter used a causal model workhorse to introduce the notion of internal validity. Figure 4.5 reintroduces that causal model to show how each type of threat to validity relates to the fundamental problem of program evaluation. The evaluator's dilemma is to isolate the impact of the program being evaluated ($X$) from that of the other two influences on the program outcome ($Y$). The other two influences are, first, other systematic factors ($Z$) that influence Y *and* that are also related to the program ($Z$), and, second, random influences ($\mu$). Both of these factors (the extraneous, confounding $Z$ variables and the random factors $\mu$) can be reinterpreted in terms of how they reflect each of the four threats to validity. The top panel of Figure 4.5 illustrates the threats to defensible causal inferences diagrammatically. The lower panel defines the statistical model that will actually be used to do all of this work. These are considerable demands that are placed on this simple statistical model, and forthcoming chapters illustrate the econometric adaptations that need to be made to address these threats.

As shown in the diagram in the top panel of Figure 4.5, internal validity concerns the accurate estimation of $\beta$, the impact of the focal program (represented as $X$), on the focal outcomes or outputs (represented as $Y$). The core of internal validity concerns our ability to separate the two systematic influences ($X$ and $Z$) on $Y$ from each other. If we cannot disentangle the relation be-

tween *X* and *Z,* then the association of *Z* with *X* will be reflected in our estimate of $\beta$, which, for internal validity, should reflect only the impact of *X* on *Y,* independent of *Z*-variables.[14]

Statistical validity concerns the accuracy with which we can separate the influence of *X* on *Y* from that of $\mu$, the random component. For example, a large sample reduces the $\mu$ component, making it easier to separate the "signal" (*X*) from the "noise" ($\mu$). Measurement (un)reliability refers to random measurement error (RME). There can be RME in treatment (*X*) or in outcome variables (*Y*). RME in outcome variables is a source of statistical invalidity; it is also represented by $\mu$. In other words, a small sample size, the main component of statistical invalidity, is one source of random error in *Y.* Random measurement error in *Y* is another source. Thus the $\mu$-term captures issues of both statistical (in)validity and measurement (un)reliability, or RME, in *Y.*

Issues of measurement also affect internal validity. For example, there can also be RME in the program variable, *X,* which we represent in the diagram by $\mu_x$. Oddly enough, but clarified by the diagram, random measurement error in program variables introduces (random) error into the *X*-variable. It poses a threat *not* to statistical validity but rather to internal validity, because its effects, if uncontrolled, are picked up by b. Similarly, NRME or (in)valid measurement of *Y* (or *X*) is also a threat to internal validity. Measurement (in)validity refers to the possibility that a nonrandom component (a *Z*-term) can slip, undetected, into the measurement of *Y* (or *X*), making it difficult to determine whether it is *X* that affects *Y* alone, or whether it is *X* along with an unintentionally invalid instrument for measuring *Y* (or *X*), that is affecting *Y.* Undetected, the invalid measurement of the outcome *Y,* or the program *X,* becomes, conceptually, a part of the *Z*-variable. The correction is to measure this systematic (measurement) error and to remove it in some way, just as we use different research designs to remove other threats to internal validity by separating the effects of *X* on *Y* from those of *Z.*

Even parts of external validity can be clarified using the causal model workhorse. Specifically, statistical interaction means that one causal model might pertain under some circumstances, while another pertains under other circumstances. For example, β might be positive for one group of people, while it could be negative for another.

We summarize these as follows:

| **To attain:** | **Try to:** |
|---|---|
| Internal validity | Separate the impact of *X* from *Z* |
| Measurement validity (same as internal validity) | Separate the impact of *X* from *Z* |
| Statistical validity | Separate the impact of *X* from $\mu$ |
| Measurement reliability of *Y* (same as statistical validity) | Separate the impact of *X* from $\mu$ |
| Measurement reliability of *X* (same as internal validity) | Separate the impact of *X* from *Z* |
| External validity (undetected statistical interaction) (same as internal validity) | Establish conditions for causal model |

The goal of program evaluation remains unchanged. It is to find out what would have happened to the outcome or output (*Y*) if the program (*X*) had not been implemented or had been implemented differently, net of the effects of the confounding influences of extraneous systematic factors (*Z*) and random effects ($\mu$). The causal model workhorse is a useful tool for clarifying the four types of threats to validity or errors of causal conclusions that can happen on the way. The model allows us to see the overlap among these four threats. Some aspects of measurement

(un)reliablity are really a threat to statistical validity; some threats to measurement (un)reliability, as well as threats to measurement and external validity, are really threats to internal validity. Thus, in the pages that follow, we focus only on the internal and statistical validity of the basic designs, because these two general types of validity subsume the other two subtypes.

## Basic Concepts

Internal validity: definition
What makes *Y* change?
- *X* Intervention variable(s)
- *Z* Confounding or extraneous variable(s)
- $\mu$ Random factors

The basic equation
The graphic of the equation
Examples of *Z*-variables as threats to internal validity
Types of comparison designs
CS designs: examples
TS designs: examples
Threats to internal validity
- History or intervening events
  - Definition
  - Examples
- Maturation, secular change, or trends
  - Definition
  - Examples
- Testing
  - Definition
  - Examples
- Instrumentation
  - Definition
  - Examples
- Regression artifacts or regression to the mean
  - Definition
  - Examples
  - (Two-way causation as one example)
- Selection or uncontrolled selection
  - Definition
  - Examples
  - (Two-way causation as one example)
- Experimental mortality or attrition
  - Definition
  - Examples
- Multiple treatment interference
  - Definition
  - Examples

- Contamination
  - Definition
  - Examples
- The relation between type of threat and type of comparison (CS vs. TS)
- Three basic research designs for reducing selection threats
  - randomized field experiment (RFE)
  - quasi experiment (QE)
  - time-series (TS)
  - cross-section (CS)
  - nonexperiment (NE)

The basic causal diagram and threats to validity

- Threats to internal validity
  - Threats due to confounds ($Z$)
  - Threats due to RME in $X$, NRME in $X$, NRME in $Y$
  - Threats due to undetected statistical interaction or low external validity
- Threats to statistical validity
  - Threats due to random components ($\mu$)
  - Threats due to RME in $Y$

## Do It Yourself

Suppose that the two causal claims below are the conclusions of actual empirical studies. Without knowing any specifics about the research studies that produced each claim, what are the potential threats to the internal validity of each causal claim? Explain *why* each one might be a potential threat.

(Example for CS design)
1. Decentralized (as opposed to centralized) World Bank projects have a higher rate of return.

(Example for TS design)
2. Putting more police out on the streets has reduced the crime rate in this neighborhood.

## A Summary of Threats to Internal Validity

Definition: internal validity = accuracy of causal claim

1. History or external events: in TS studies, an event other than the change in the treatment ($X$) might have caused the outcome ($Y$) to change (or might cause some of the observed net change in $Y$).
2. Maturation, trend, endogenous change, or secular drift: in TS studies, $Y$ may be changing partly or entirely because of an underlying trend and not because of change in the treatment ($X$).
3. Testing: in TS studies with obtrusive measures, taking the test, and not change in the treatment ($X$), may cause the outcome ($Y$) to change (or might cause some of the observed net change in $Y$).

4. Instrumentation: in TS or CS studies, change in the calibration of the measurement procedure or instrument, rather than the treatment (*X*), may partly or entirely cause the outcome (*Y*) to change.
5. Regression artifacts: in TS or CS studies, when subjects are selected for treatment because of extreme scores, an observed change or difference in outcome (*Y*) may be observed partly or entirely because of the random tendency of extreme scores to return to their normal value.
6. Selection: in XS studies, when the groups to be compared differ on factors besides the treatment (*X*), then these differences (*Z*) may account partly or entirely for the observed differences in outcome (*Y*).
7. Experimental mortality or attrition: in TS or CS studies, when two or more groups are being compared, observed between-treatment differences in outcome (*Y*) may be partly or entirely attributable to a differential loss of respondents rather than to the treatment (*X*).
8. Multiple treatment interference: in TS or CS studies, when one treatment (*X*1) is confounded with another (*X*2), then it is impossible to separate the impacts of one treatment from the other.
9. Contamination: in CS studies, when the separation of treatment groups from control groups is less than it should be, or when the control group receives some of the treatment, then the overlap may partly or entirely account for the observed between-treatment difference (or lack of difference) in outcomes (*Y*).