CHAPTER 16

Measure Your Mission

What You'll Learn

- It is important that you measure outcomes and treatments that correspond to your mission.
- If you measure the outcome in an incomplete way, apparent improvements may be misleading.
- Data always comes from a particular context. When applying the lessons drawn from data to a new context, it is important to think clearly about whether the contexts are sufficiently similar that the lessons will continue to hold.
- Sometimes, there is a relationship in the world that would help you achieve your goal. But once you actually use that relationship to try to do so, the relationship itself disappears, so it is no longer helpful.

Introduction

When you use evidence to inform your decisions, you have some goal in mind. That goal is your mission. Why is it important to measure it?

Suppose you have evidence about a causal relationship; some action affects some outcome in a predictable way. If changing that outcome means you have achieved your goal—that is, if in measuring the outcome you measured your mission—then knowledge of that causal relationship is straightforwardly useful. But what if changing the outcome you measured doesn't necessarily mean you have achieved your goal, or what if it only corresponds to one part of your goal? Then, which action the evidence suggests will further your mission might not be so clear.

The same goes for correlations. Suppose your mission involves trying to predict some outcome, but you've measured a related, though different, outcome. Are you sure that the correlates of the outcome you measured will help you predict the outcome of interest?

In this chapter, we will explore several ways in which things can go wrong when we have good evidence about what might turn out to be the wrong thing. Each of these examples will illustrate the reasons it is important to measure your mission, as best as possible, when trying to use evidence to make better decisions.

Measuring the Wrong Outcome or Treatment

The most straightforward way that you might fail to measure your mission is by measuring an outcome or treatment that doesn't quite correspond to what you are really interested in. Here we consider three ways in which this commonly happens.

Partial Measures

Often our mission is to change some outcome—say, educational achievement, national security, or health—that is hard to measure in its entirety. For instance, we might not have an encompassing measure of overall educational achievement, but perhaps we can measure whether standardized test scores improve. Such partial measures can be helpful. But we have to be careful about interpretation because improving test scores is not our mission. Our mission is improving education.

In many settings, there are good reasons to think that improvements on one dimension might tend to coincide with losses on other dimensions. That is, as we get better at one part of a problem, we might get worse at other parts. A simple reason for this is resource constraints. Suppose your overall mission is to make a local park more beautiful. You have a budget to support your mission. If you spend more resources on trash pickup, you have less money to spend on landscaping. So improving on one dimension means getting worse on another. And if you just have a partial measure of your mission (say, the amount of trash on the ground), then as you spend more money on trash pickup, you might be tempted to conclude you are doing a better job achieving your mission. But, because things are getting worse on the landscaping dimension as a result of devoting more resources to trash pickup, this is misleading.

There are additional reasons, besides limited resources, for a negative correlation across dimensions of a problem. Perhaps the most interesting is *strategic adaptation*—efforts to improve outcomes on some dimension lead people to adjust their behavior to get around those efforts. This too can make partial measurements problematic. Let's see how this plays out in an example.

Metal detectors in airports

Starting in the mid-1960s, hijacking became a serious problem in U.S. civil aviation. Over eighty airplanes were taken by hijackers in 1969 alone. The hijackers included Americans, Croatians, Cubans, Japanese, North Koreans, Palestinians, and many others. Their motivations ranged from simple ransom to nationalist, leftist, and other global political causes. In the early 1970s, in response to this growing threat to air safety, the United States increased airport security. Most importantly, metal detectors were installed in every major U.S. airport in early 1973.

Imagine you were a government official tasked with evaluating the efficacy of these heightened security measures. A natural question you might ask is whether they resulted in a significant decrease in hijackings. Figure 16.1, showing hijackings per quarter from 1968 to 1978, suggests the answer is yes. Prior to 1973 (represented by the dashed, vertical line), there was an average of almost twenty hijackings per quarter. But after 1973, that number drops to fewer than ten per quarter.

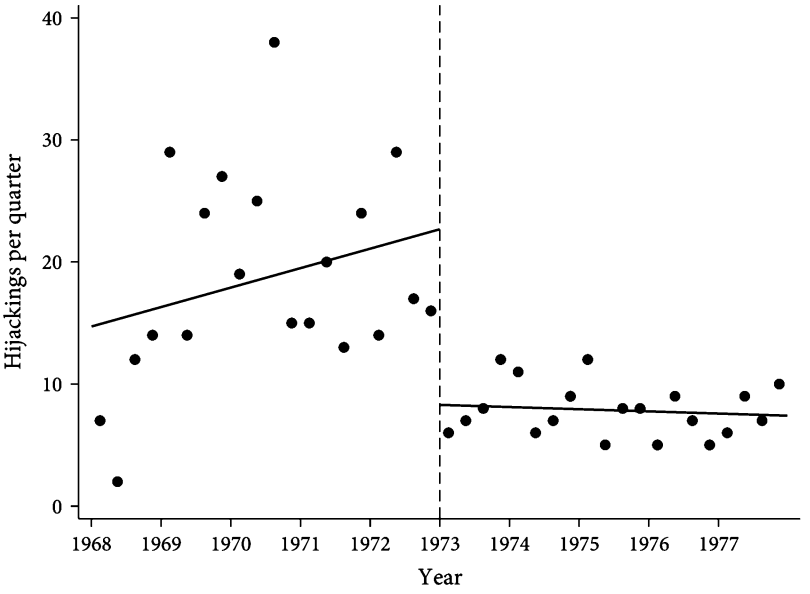


Figure 16.1. Quarterly hijackings 1968–1978 along with separate regression lines for quarters before and after first-quarter 1973. The vertical dashed line indicates when metal detectors were installed in U.S. airports.

Let's think about whether we've measured our mission. One possibility is that the mission is to reduce hijackings. In that case, hijackings are the right outcome to study and this looks like a success. But another possibility is that the mission is to increase security from all terrorist hostage takings, not just hijackings. In that case, hijackings are only a partial measure of the mission because there are lots of other kinds of terrorism.

Moreover, this is just the kind of setting where we might worry that improvements on one dimension of a problem (here, hijackings) tend to coincide with exacerbation of the other dimensions of the problem (here, other kinds of terrorist attacks). The reason is strategic adaptation. As airport security improves, we might worry that terrorists substitute hijacking for other kinds of hostage takings. If this is the case, the apparent reduction in hijackings might be misleading as a measure of how successful increased airport security was in terms of the overall counterterrorism mission.

And, indeed, this appears to be the case. Figure 16.2 shows a finding inspired by the work of Walter Enders and Todd Sandler—after metal detectors were installed in U.S. airports, other kinds of terrorist hostage takings became more frequent. And so, if we have a more encompassing, rather than partial, measure of our mission, we reach somewhat different conclusions.

Of course, this doesn't mean that the metal detector policy was a failure. The substitution from hijackings to other hostage takings does not appear to be one-for-one. Moreover, hijackings might be worse, on average, than other kinds of hostage takings. So this might still be a counterterrorism win. But it is not nearly so dramatic a win as one might have thought looking only at the impact on hijackings rather than a more complete measure of the mission.

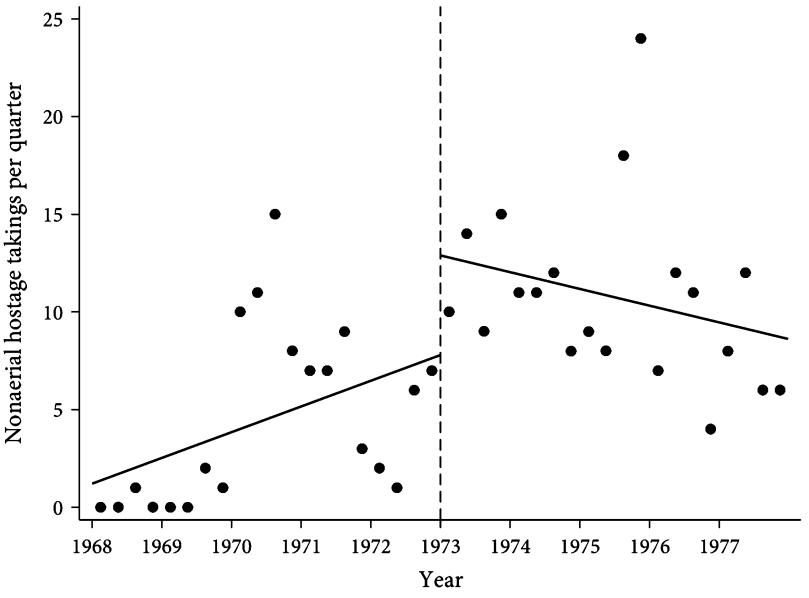


Figure 16.2. Quarterly hostage takings not on airplanes, 1968–1978, along with separate regression lines for quarters before and after first-quarter 1973. The vertical dashed line indicates when metal detectors were installed in U.S. airports.

Intermediate Outcomes

Often, measuring the outcome associated with your mission is difficult, takes a long time, or just doesn't yield enough data. A common solution is to measure intermediate outcomes, steps along the path of the mission, that we hope will be indicative of the longer-term objective.

Suppose you're running a political campaign and you're trying to maximize the probability that your candidate wins. You want to test a few different ads to see which ones are the most effective. You could run the ads in different media markets and see where you do the best on election day. But that won't do much good. You need to know which ad to run before the election happens. So, instead, you have to measure some outcome that will give you some sense of which ad is best while the campaign is ongoing.

One natural option to help you decide on your strategy is opinion polls. Instead of running ads randomly and seeing what happens to vote totals, you could run ads randomly, conduct opinion polls, and see which ad appears to help you in the polls. There's not necessarily anything wrong with doing this. It's a good idea. But you have to keep in mind that you don't care about polls per se. You care about votes. So to the extent that changing poll numbers is indicative of a step along the path to changing votes, learning about this intermediate outcome is informative about your mission. But, for example, it's possible that your ad changes poll numbers by changing which people are willing to respond to the poll or whether people tell the truth to pollsters, without changing people's actual vote choices. If that is the case, impacting the intermediate outcome might not matter at all for the final outcome you care about. So, whenever you use an intermediate outcome instead of a measure of your ultimate mission, you want to think about

how sure you are that the intermediate outcome really is a step on the path to your actual goals.

Let's think about an example from medicine, where the impracticality of studying the actual outcome of interest is often particularly acute.

Blood pressure and heart attacks

Suppose the goal of a new drug is to reduce heart attacks. Unfortunately for research purposes (but fortunately for other reasons) heart attacks are rare. So relatively few people in any given sample will have heart attacks during the course of a drug trial. As such, it is very hard to learn directly about whether a drug reduces heart attacks even in a well-designed experiment that randomizes who gets the drug and who doesn't.

So what do medical researchers do? One alternative to waiting twenty years to see whether patients who were assigned the drug are less likely to have heart attacks is to study an intermediate or surrogate outcome, like blood pressure. Since blood pressure predicts heart attacks, the thinking goes, if a drug reduces blood pressure, it is likely to reduce heart attacks.

But we have to be careful. We learned in part 2 that correlation need not imply causation. Playing basketball is correlated with height but experimentally increasing basketball playing does not increase height. Similarly, just because blood pressure and heart attacks are correlated doesn't mean that a drug that reduces blood pressure will reduce heart attacks. For that, you'd have to have compelling evidence that blood pressure has a causal effect on heart attacks.

Now, there are good reasons to believe that blood pressure really does have a causal effect on heart attacks. But for many other intermediate outcomes used in medical studies, the causal linkage may be less clear.

In a 1994 review of the evidence, Thomas Fleming illustrates the point in a discussion of research on cancer. Often, when studying cancer treatments, scientists cannot wait long enough to look at the effect of a treatment on, say, mortality. So, instead, they study the effect on an intermediate outcome. One popular such intermediate outcome is tumor size.

For instance, Fleming describes a medical trial for a drug intended to treat prostate cancer. The researchers determined that if they examined mortality as their outcome, they would need a sample of between 40,000 and 100,000 subjects to detect a reasonably sized effect because death from prostate cancer is rare and slow. Since they could only recruit 18,000 men for their trial, they instead decided to use tumor size, as measured by a prostate biopsy, to assess the effectiveness of the drug.

One problem, as Fleming discusses, is that prostate tumor size is only a very weak proxy for the actual mission, which is presumably not dying from cancer. Thirty percent of men over the age of fifty test positive for prostate tumors. But only 3 percent actually die from prostate cancer. Many prostate tumors grow very slowly. So other things, like heart attacks, get people first. The experiment showed that the drug being tested significantly reduced tumor size. But it is entirely possible that much of the reduction in tumor size was in the kinds of tumors that were never going to harm the subjects in the first place. So we really don't know whether progress on this intermediate outcome contributed much, if anything, to progress on the mission of avoiding death by prostate cancer.

Of course, we don't mean to suggest that studying intermediate outcomes is a bad idea. Indeed, it is often the best that can be done, given other constraints. But in

interpreting the finding of a relationship between some action and an intermediate outcome, it is important that we think clearly about what we know about the relationship between the intermediate outcome and our actual mission.

Ill-Defined Missions

Often, your mission may be slightly tricky to pin down. In particular, there is sometimes more than one reasonable way to measure what may seem to be the same mission. But which choice you make can matter a lot. So it is important to think hard about what outcomes and treatments really define your mission.

Suppose you're a college student considering your educational and career choices with the goal of maximizing your future earnings. The first thing you might think to do is study the Forbes list of the richest people in the world and try to follow in their footsteps. One thing you might infer is that the way to maximize your earnings is to drop out of college and start a tech company. This was the strategy taken by Bill Gates, Mark Zuckerberg, and Larry Ellison, three of the eight richest people in the world at the time of this writing. But you won't make that mistake because you learned in chapter 4 that correlation requires variation. To know if dropping out of college and starting a tech company is correlated with success, you can't just study the most successful people.

Suppose you pushed further and tried to get a sense of how many people in the underlying population dropped out of college and started their own tech company. You'd surely find that less than .01 percent of all people dropped out of college and started their own tech company, and yet 37.5 percent of the world's eight richest people did so. So there appears to be a strong correlation. People who drop out of college and start their own tech company are much more likely to end up one of the world's eight richest people than people who stay in college or never start a tech company.

Having identified a correlation, there are still some reasons you might not want to make a rash decision and drop out of college today. First, we might have just inadvertently engaged in something akin to *p*-hacking. We studied a small population of extremely wealthy individuals, we looked for commonalities, and we eventually found something that a few of them have in common. But that might just be a coincidence. Maybe the correlation we observe today won't hold in the future, in which case dropping out and starting a tech company might be a bad idea.

Yet another reason we wouldn't recommend dropping out and starting a tech company is that we're not comparing apples to apples. The kinds of people who drop out of college and start a tech company are likely different from those who don't, and we have little way of knowing whether they would have been equally successful had they not dropped out. That is, following the lessons of chapter 9, this correlation is not an unbiased estimate of the causal relationship.

But even setting all of these reasons aside, there's a fundamental problem with this line of thinking that has to do with correctly measuring your mission. What outcome do you really care about? Is it your expected earnings or is it your probability of becoming a multi-billionaire? To the extent that dropping out of college and starting your own tech company makes you more likely to be one of the richest people in the world, it probably also makes you more likely to be in serious debt. And for all we know, it might significantly reduce your expected earnings, even if it increases your chances of becoming very wealthy. Is that a gamble you're willing to make?

We're not here to tell you what your particular objectives should be. Some people may have a deep desire to become a billionaire, which makes them willing to take significant

risks. But we suspect most people are more averse to risk and would rather maximize their expected earnings or perhaps even minimize their chances of being in poverty. Your particular objective should inform the analyses you conduct. If your goal is to maximize expected earnings, it might be a huge mistake to examine the correlates of being on the Forbes list of wealthiest individuals. Instead, you'd want to collect data on earnings to see how various educational and career choices correspond with earnings, on average. We suspect you'd find that graduating from college and perhaps even going to professional school is a better predictor of earnings than dropping out and starting your own company.

This mistake of studying the wrong outcome can be made in the other direction as well. If you're managing a political campaign or coaching a sports team, you don't really care per se about your expected point margin or vote share. What you care about is winning, so you should choose strategies that maximize that objective. For example, if your political candidate is polling badly and there's only a week left in the campaign, you might be willing to gamble on an otherwise ill-advised strategy to give yourself a chance of winning. Maybe you decide to roll out a really aggressive new policy proposal that the voters probably won't like. In expectation, such a strategy reduces your vote total. But there is a small chance the voters will love your wild idea and you will win. If you don't really care about vote share (losing by five points or ten points is still losing) but only care about winning, a strategy that hurts your expected vote share may be optimal.

And, of course, this measurement problem doesn't only apply to outcomes. It also applies to measuring treatments. This is perhaps most clear when the variables we measure are meant to represent abstract concepts. We have to think clearly about what, exactly, we are measuring when we rank some countries as more or less democratic than others or some classes more or less difficult than others. But this concern can also emerge even when we are measuring more concrete quantities in the world. Here's an example.

Climate change and economic productivity

Many people are interested in the long-run effects of climate change on economic growth. Climate change, of course, happens over a long period of time and, thus, is hard to measure and study. But related phenomena, such as weather and temperature, vary frequently. So scholars sometimes use variation in the weather to try to learn about the effects of climate change.

For instance, Marshall Burke, Solomon Hsiang, and Edward Miguel estimate the effect of unexpected temperature fluctuations on GDP growth using a difference-in-differences design. That is, they compare the GDP within a country in years when it is exposed to warmer- versus cooler-than-average temperatures due to naturally occurring atmospheric variation. They find that economic productivity is maximized at an annual average temperature of 13 degrees Celcius and that it declines precipitously as the temperature rises. They conclude that "if future adaptation mimics past adaptation, unmitigated warming is expected to reshape the global economy by reducing average global incomes roughly 23% by 2100."

This is an important study and an important conclusion. But the caveat offered by the authors, "if future adaptation mimics past adaptation," points to a critical measurement issue.

The authors are interested in the effects of *climate change*. But the treatment they measure is *temperature fluctuations*. Climate change happens slowly, giving people and society time to adapt. Temperature fluctuations happen quickly, making adaptation difficult. Moreover, unlike temperature fluctuations, climate change is associated with shifts in weather variability, disease vectors, natural disaster prevalence, and so on. Thus, in important ways, temperature fluctuations do not measure the right treatment. And, in particular, they don't measure the right treatment in ways that are relevant for the question of productivity. In light of these measurement concerns, we probably shouldn't put a lot of faith in that 23 percent estimated effect.

To appreciate the distinction, consider the difference between the effects of a hot day on economic productivity versus the effects of a hot century. We live in Chicago, which can be a pretty cold place. If we were pleasantly surprised by an especially warm day, Anthony might be tempted to leave work early to play golf. But if climate change meant that every day was warmer, he wouldn't quit his job and play golf every day. And if it meant that days were warmer, but storms were more frequent, who knows what would happen to his golf playing. The fact that unexpected hot days decrease productivity does not necessarily tell us the long-run effects of climate change because we haven't measured and studied the right thing.

Do You Have the Right Sample?

Studying the right outcome and the right treatment isn't all there is to measuring your mission. We also need to make sure we have the right sample.

When applying evidence to decision making, we almost always have to take knowledge gleaned in some place and time and try to apply that knowledge to understand what will happen in another place and time. Essentially we are making an analogy between the contexts in which the evidence was generated and the contexts in which we now wish to apply the lessons we learned from that evidence. So we always have to ask whether those contexts are sufficiently similar that such an analogy is valid. Otherwise, we may take actions that are consistent with achieving our mission, but only in a very different context from the one in which we are acting.

External Validity

The basic problem here is that relationships can differ from context to context. We've spent a lot of time so far in this book on what is sometimes called *internal validity*. Internal validity is about credibly estimating the estimand (e.g., Is the estimator unbiased?). But even if you've done everything right with respect to internal validity, you still need to be able to think clearly about whether that relationship is likely to also exist in the context where you hope to apply it. Broadly, this is the problem of *external validity*. External validity is about whether there are good reasons to believe that a relationship estimated on data from one context will hold in some other context. An example will help to illustrate the point.

Malnutrition in India and Bangladesh

In the 1980s, the World Bank implemented the Tamil Nadu Integrated Nutrition Project (TINP) in a region of southern India where malnourishment was endemic. While the

project included some resources for supplementary nutrition, the main focus was on helping mothers, the main household decision makers concerning food purchasing and preparation, make better use of the resources already at their disposal. The TINP is viewed as a major success by the World Bank. And, while there is some debate, it is widely credited with making a major difference in reducing malnourishment and malnutrition in Tamil Nadu.

This apparent success inspired the Bangladesh Integrated Nutrition Project (BINP) in the 1990s. By that time, Bangladesh, which borders India to the east, was among the most malnourished countries on earth. Evidence suggests that, in the early 1990s, almost two-thirds of Bangladeshi children under the age of five had growth stunting due to malnourishment.

Because the TINP had been rigorously evaluated and shown to have made a significant and meaningful dent in malnutrition, the BINP was modeled quite directly on the TINP. And so scholars and practitioners alike were surprised when the BINP's impact did not live up to the promise of the TINP. Despite being designed to replicate perhaps the most successful malnutrition intervention in history, rigorous evaluation shows little to no impact of the BINP on malnutrition. What went wrong?

There are, of course, many possible answers. And it is virtually impossible to know for sure why the program failed. But one important factor seems to have been a cultural difference between Tamil Nadu and Bangladesh. As we've mentioned, in Tamil Nadu, mothers are typically the chief decision makers regarding food purchasing and preparation. Thus, it made sense to target mothers for the TINP's nutritional education efforts.

This focus on mothers was exported directly from the TINP to the BINP. But in many households in Bangladesh the father or the mother-in-law (i.e., the father's mother), rather than the mother, has authority over food purchasing or preparation. Because this was not the case in Tamil Nadu, these important decision makers were not targeted by the BINP. Thus, the BINP may have failed, at least in part, because a targeting decision that made perfect sense in one setting was no longer so sensible in another.

This example is particularly interesting to us because it points to the potential for a complementarity between quantitative evidence and qualitative knowledge. Assessing the impact of the TINP required a quantitative approach. But attempting to apply that knowledge to the Bangladeshi context went wrong because of a lack of knowledge about key cultural and institutional differences between Tamil Nadu and Bangladesh. A team that combined both people with expertise in quantitative assessment who could think clearly about the causal effect of the TINP and people with deep qualitative knowledge of the two contexts might have resulted in a better outcome than either alone could hope to achieve.

Selected Samples

A particularly common way for people to end up measuring their mission in the wrong context is by studying selected samples. A *selected sample* is a sample of observations that wasn't drawn at random from the population of interest but rather was selected to be studied because it possessed some set of characteristics. The problem, of course, is that a selected sample may not be representative of the population as a whole. And relationships that hold in that selected sample may not hold in the broader population. If your mission is to predict, understand, or influence the behavior of the broader population, things can really go wrong if you rely on evidence from a selected sample.

College admissions

Here's an example that is near and dear to our hearts. Standardized test scores, for better or worse, have been an important part of the college admissions process for decades. However, in 2018 our own university announced that it would no longer require applicants to submit such scores. (Several other colleges and universities have done likewise.) One (among several) of the rationales for going test-optional was evidence-based. University leaders looked at the students who attended the university and found little correlation between test scores and performance. So, the argument went, maybe test scores aren't very good predictors of college performance.

The mission of a college admissions office is multifaceted. But part of that mission is to identify the most academically talented students from the pool of applicants. To fulfill this mission, the admissions office would like to know whether some characteristic of *applicants* (here, their test scores) is correlated with academic performance in college. But that is not the question the exercise described above addresses. Rather, that analysis asks whether some characteristic of *enrolled students* (namely, their test scores) is correlated with academic performance in college. But the answer to those two questions need not be the same.

The set of enrolled students is a selected sample of the set of applicants. Students were admitted to college based on their test scores and other factors like writing ability, teacher recommendations, grades, community service, and overcoming adversity. The fact that test scores were used in admissions can lead to a fundamentally different correlation between test scores and academic performance in the selected sample of enrolled students and the broader set of applicants.

To see how, think about students with low test scores who were nonetheless admitted to the university. Those students must have had some other characteristics that led the admissions office to overlook their low scores. Maybe they wrote stellar essays, had particularly strong recommendations from teachers, or made great grades in high school. Similarly, students with particularly high test scores were likely to have been admitted even with somewhat weaker performance on these other dimensions. For this reason, in the set of admitted and enrolled students, we might expect a negative correlation between test scores and other markers of academic quality.

Now, it's quite plausible that test scores are a good predictor of college performance among all applicants, but that writing ability, teacher recommendations, and high school grades are also good predictors. Therefore, once we look at the selected sample of enrolled students, we'll find a weak correlation between test scores and performance. But that's because the only people with low scores who got in are people who are really strong on other dimensions. So that weak (or non-existent) correlation in the selected sample of enrolled students does not mean that test scores are a bad predictor of academic performance among applicants.

This issue of studying selected samples is also prevalent outside the context of college admissions. So let's talk through one more example: baseball.

Why can't major league pitchers hit?

Major League Baseball fans know that pitchers tend to be the worst hitters on their teams. In the National League, where pitchers are required to bat, managers will typically have their pitchers hit last in the lineup to minimize their trips to the plate. And if a pitcher is coming out of the game, the manager will always replace them with a

pinch hitter. In the 2017 Major League Baseball season, the batting average for the average pitcher was .125. The batting average for the average non-pitcher was .259. This is a massive difference. The American League has a designated hitter rule specifically so that pitchers don't have to bat.

So why are major league pitchers so bad at hitting? If you ask a baseball expert, they'd probably tell you that pitchers spend so much time practicing their pitching that they don't have time to practice their hitting. And there might even be something about great pitchers that makes them weaker hitters. Perhaps the kind of strength, flexibility, or body type that's good for pitching is bad for hitting.

These explanations sound pretty compelling, and they're probably right to some extent. But they also probably are not the whole story. One way to start to see this is to notice that this pattern does not hold for high school baseball.

We collected data on four Chicago-area high school baseball teams from the 2018 season and calculated the batting average for non-pitchers and pitchers (defined as players who pitched more than ten innings in the season). Unlike the pros, among these high school players, the pitchers actually have slightly higher batting averages than the non-pitchers: .322 versus .317.

How can this be? Why is the correlation between pitching and hitting ability slightly positive for high school baseball players but negative for seasoned professionals? It's not as if pitching practice doesn't crowd out batting practice for young pitchers. And you'd think the arguments about physical specialization would apply in high school as well as in the major leagues. So why does the correlation seem to change so dramatically and even flip signs as players age?

Even at the professional level, we can see that there wasn't always a negative correlation between pitching and batting ability. Figure 16.3 shows the batting average for the average pitcher and the average non-pitcher in Major League Baseball from 1871 to 2017. In the nineteenth century, pitchers and other position players had comparable batting averages. But starting in the twentieth century, the pitchers appear to get worse at hitting relative to other players, with the gap gradually increasing over time. And in the modern era, as we already discussed, non-pitchers get approximately twice as many base hits per at bat as do pitchers.

We suspect that the explanations for the changing correlation over time and the difference in correlation between high schoolers and the professionals is one and the same. And it has to do with selected samples.

Start by thinking about the correlation between pitching and batting ability in the entire population. Suppose we just randomly sampled individuals (say, teenagers and older) from the whole world and asked them to play baseball so we could measure their pitching and batting abilities. What do you think we would find? We suspect that we'd find a pretty strong positive correlation. Some people are athletic and have experience playing baseball. They are likely to be good at hitting and pitching. Other people are uncoordinated and inexperienced. They are likely to be bad at hitting and pitching. So, in the population as a whole, you're likely to find exactly the opposite correlation from what you find among professionals.

To see why this is, think about how a person becomes a Major League Baseball player. They almost surely play high school ball. A high school coach is trying to assemble the best team possible. That involves choosing the players, from the set of players available, who offer the best combination of batting and pitching ability. To make your high school team, you have to be pretty good at some combination of hitting and pitching. But you don't have to be amazing—you can be a good hitter (even if you are a bad

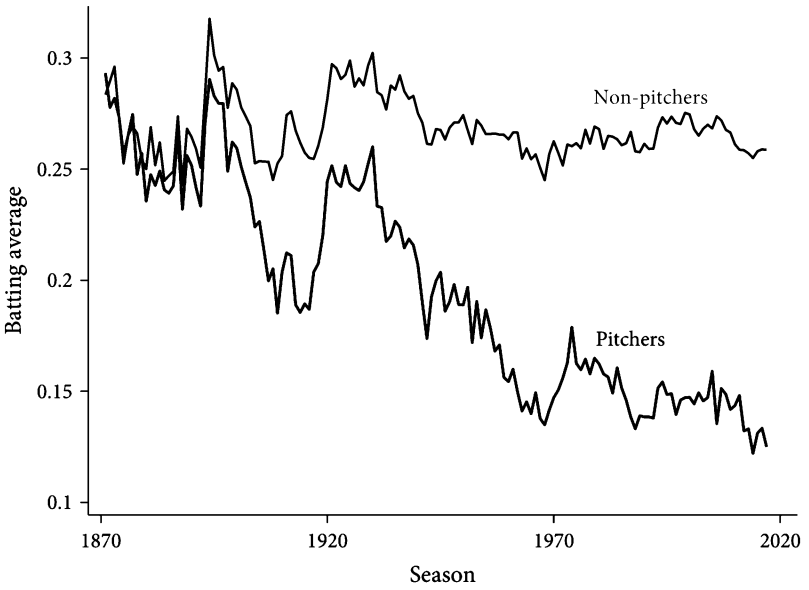


Figure 16.3. The negative correlation between pitching and hitting skill in the major leagues has gotten stronger over time.

pitcher), a good pitcher (even if you are a bad hitter), or an at least passable pitcher and hitter.

For most players, the path to the major leagues runs through the minor leagues. Minor league coaches are also trying to assemble the best team possible. And so, they too need the best combination of hitting and pitching available. So, to make a minor league team, you have to be really, really good at some combination of hitting and pitching. That means being a great hitter (even if you are a bad pitcher), a great pitcher (even if you are a bad hitter), or an at least good pitcher and solid hitter.

Finally, to make the major leagues (at least the National League, where pitchers hit), the test is even more stringent. You've got to be a truly amazing hitter (even if you are a bad pitcher), a truly amazing pitcher (even if you are a bad hitter), or a pretty great pitcher who can also hit.

Below is a simple demonstration (with hypothetical data) of what these ever more stringent selection criteria do to the correlation between hitting and pitching ability in different samples. Suppose (for simplicity) that we can give every potential baseball player a score that separately summarizes their pitching and batting abilities, and teams want to recruit players with the highest possible sum of both pitching and batting ability. How high that sum needs to be is increasing as you move up the ranks of baseball.

In the top-left panel of figure 16.4, we've drawn a scatter plot of some data with a strong positive correlation between pitching (horizontal axis) and batting ability (vertical axis). This is meant to represent the entire population. If we just let everyone on the team (as an entry-level team for kids might do), we'd see a pretty strong positive correlation between pitching ability and batting ability. This seems right to us. Our memory of youth sports is that the kids who were the best at one aspect of the game were often the best at every aspect of the game.

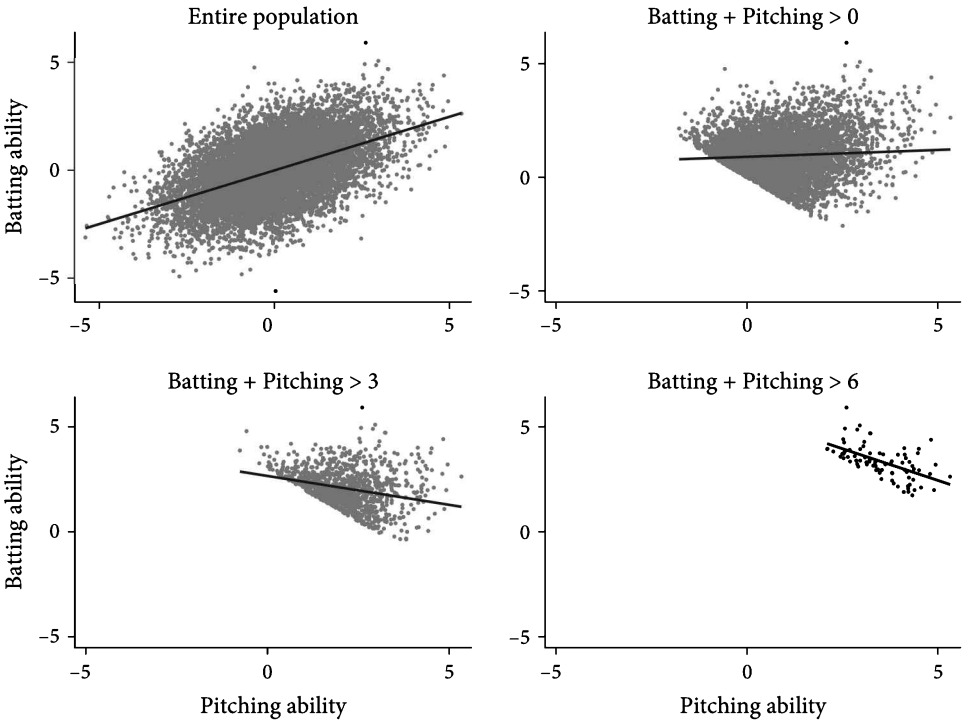


Figure 16.4. Increased selection can turn a positive correlation into a negative correlation.

The upper-right panel of figure 16.4 is meant to represent high school or the major leagues in the nineteenth century. In high school, the coach is only willing to accept players who are above average in the population, which translates into having a sum of batting and pitching ability above 0. Similarly, in the nineteenth century, baseball wasn't that popular, so professional coaches couldn't be so selective. You can make a team like this by being a good batter (say, a 3) and a weak pitcher (say, a −2); a good pitcher (say, a 3) and a weak hitter (say, a −2); or a slightly above-average pitcher and hitter (say .5 on both). But this level of selectivity eliminates people who are bad at both. And so, among the selected sample of high school or nineteenth-century baseball players, there's little correlation between batting and pitching ability.

The panel on the bottom left might represent the minor leagues or the early twentieth century. Selectivity has increased. Players need at least a 3 to make the team. So bad pitchers must be genuinely good hitters, bad hitters must be genuinely good pitchers, and some players can make the team by being solid at both. By eliminating even more players who are only okay at both, this level of selectivity flips the relationship from that in the population—inducing a slight negative correlation between batting and pitching ability.

Finally, the panel on the bottom right might represent the contemporary National League. Here selectivity is very high, since only an elite few can make it. So bad pitchers must be truly great hitters, bad hitters must be truly great pitchers, and for players to make it with a combination of skills, they must be terrific at both. At the highest levels, then, we expect to see a strong negative correlation between pitching and batting ability, even without any difference in time spent practicing or body type.

Figure 16.4 illustrates a fairly general phenomenon. Correlations in selected samples are often quite different from those in the broader population. This is important because we often only have data on selected samples. But we may want to make predictions and inferences about the broader population.

If you're a baseball scout, you have rich data on major league players available to you. You might try to use the data to predict who will become a star player so you can search for those features among high school and college athletes. But if you look for correlations of great performance among major league players, you'll draw misleading inferences. For example, you might find that slow runners are really good power hitters. So should you go find the slowest runners and recruit them to play professional baseball? Surely not! The reason that slow runners are good power hitters in your selected sample is the same as the reason good pitchers are bad hitters. The only way a slow runner can make it to the major leagues is if they're a great power hitter.

Strategic Adaptation and Changing Relationships

There is another key issue for thinking about measuring your mission. Sometimes, there is a true relationship in the world that would help you achieve your goal. But once you actually use that relationship to try to do so, strategic adaption makes that relationship itself disappear or change, so that it is no longer so helpful. To see how this works, let's look at some examples of this phenomenon throughout history.

The Duty on Lights and Windows

In 1696 King William III of England needed money. Kings, of course, always need money. But this need was particularly pressing. Up until the 1660s, Britain produced coins made of hammered silver. These coins had a serious problem—people shaved or clipped the valuable silver around the edges. As a result, the value of the coins in silver was less than their face value. This widespread practice of coin clipping threatened to undermine the credibility of English currency.

To address the problem, the Crown undertook the great recoinage of 1696, offering to buy back clipped coins in exchange for new, machined coins that could not be clipped.[1] But buying back clipped coins for proper coins was expensive. The Crown essentially had to make up the difference between the face value and the value of the silver. So the Crown needed to raise revenue. But how to do so?

The Crown wanted to tax the wealthy more than the poor. One natural way to achieve this is through an income tax. But the English were opposed to income taxes because assessing income involved an invasion of personal privacy. So the Crown needed to find a way to tax wealth that was more politically palatable. The solution they landed on was a duty on lights and windows, better known as the window tax.

The window tax had the virtue that it could be assessed from outside of a home, thereby limiting any invasion of privacy. In the earliest version of the tax, the Crown established a two shilling base fee for all homes. In addition, homes that had between ten and twenty windows paid an extra four to six shillings and homes that had over twenty windows paid an extra eight to ten shillings. Windows in rooms used for work

[1] Fun fact: The new coins couldn't be clipped because they had milled edges, a feature that persists today, even though our coins are not made of precious metals. Milled edges were an innovation created by Isaac Newton in his role as warden of the Royal Mint at the time of the great recoinage.

didn't count. The exact thresholds and fees changed over time (such taxes lasted for well over a century), but you get the basic idea.

The argument for the window tax was an obvious correlation between windows and wealth (of course, they wouldn't have put it in those terms). On average, people whose houses had more windows were wealthier. Thus, by taxing windows, the Crown could raise revenues in ways that put more of the burden on the rich and less on the poor, which was the mission.

But the story doesn't end there. The English, like many others, don't like to pay taxes. And so they strategically adapted. In the short run, windows were boarded up or bricked over to reduce the taxes owed. Over the long run, architecture changed. Larger homes began to include fewer windows and more rooms that could be presented as work rooms. And so, with the passage of time, both the revenue from and the progressivity of the window tax declined.

In this case, the Crown's mission was to raise revenue in a progressive way. To do so, it needed to identify and tax wealthier people without invading their privacy. It noticed a correlation between windows and wealth, which seemed like just what it needed to achieve its mission. But, using that correlation in service of its mission led homeowners to strategically adapt their behavior such that the correlation no longer held (or, at least, held much less strongly), undermining the mission. Thus, in considering some change in behavior or policy in response to some piece of evidence, one must always ask whether the relationship uncovered by the evidence will persist once you change your behavior or policy.

The Shift in Baseball

We know we've already spent a good bit of time on baseball in this chapter. But, if you'll indulge us, we would like to do one more example. It is a pretty good one for illustrating the idea of strategic adaption changing the usefulness of a statistical relationship.

There was a time when defenders in baseball stood in their spot and waited to see if the ball would come their way. To be sure, fielders would adjust their position a bit depending on whether a left-handed or right-handed batter was up. But, for the most part, defensive strategy wasn't too complicated.

That time came to an end with the rise of big data in professional sports. Now teams have detailed spray charts for every batter. These charts provide data on how frequently each batter hits to various parts of the field, whether they hit the ball on the ground or in the air, the angle at which they make contact with the ball, and so on. Using this kind of information, teams can make well-informed forecasts about exactly where a given batter is likely to hit the ball. And armed with such forecasts, teams have started adjusting their defensive setups aggressively, batter by batter.

The most famous version of this change in defensive strategy is called the *shift*. Examining spray charts, teams discovered that when batters (especially power hitters) hit the ball on the ground, it is almost never to the so-called opposite field (for right-handed batters, this is to their right, and for left-handed batters, this is to their left). Rather, if they are going to hit the ball on the ground, they pull the ball (i.e., right-handed batters hit it to their left, and left-handed batters hit it to their right). The shift is the obvious response to this correlation—when facing a right-handed batter, shift the infield way over to the batter's left, and when facing a left-handed batter, shift the infield way over the batter's right. The benefit of such a move is that it makes it much less likely

that a ground ball that is pulled will sneak through a hole in the infield for a base hit. The cost of this strategy is that it leaves a big hole in the infield in the opposite field. But since batters find it very hard to hit ground balls to the opposite field, this cost is small.

A few teams started shifting aggressively in the late 2000s. In 2010, the Tampa Bay Rays—led by manager Joe Maddon, an early advocate for evidence-based defense—accounted for 10 percent of all shifts, although they were just one of thirty teams. Maddon consulted spray charts and strategically placed his infielders in locations that were optimal for the particular pattern of ground balls associated with each batter. The Rays and other early adopters had a lot of success. That is, there was a negative correlation between using the shift and runs allowed.

Observing this correlation, all teams started implementing the shift. In 2011, there were only about 2,000 shifts used in total across all Major League Baseball games. By 2014, that number had grown to 13,000. And in 2016, it reached over 28,000.

But something else happened too. At first, the correlation that inspired this surge in shifting held. Teams that shifted allowed fewer runs. But batters noticed that the shift was hurting them. And they strategically adapted to avoid hitting so many ground balls into the shift. Instead, they hit more balls to the opposite side of the field, and they hit more balls in the air—over the shifted infield.

As things stand today, major league teams still use the shift a lot. But, because hitters adapted, the correlation between shifting and runs allowed that drove teams to use the shift in the first place does not hold nearly as strongly as it used to. Setting defensive strategy in response to the correlation led to changes in behavior that undid that correlation. It is perhaps worth noting that Joe Maddon—the early innovator who, as manager of the Rays, helped make the shift so popular—remains a believer in evidence-based defense. He later won a World Series as the manager of the Chicago Cubs, where he employed the shift less than any other manager in Major League Baseball.

The War on Drugs

Before leaving the topic of whether things will change once you act, it is worth pausing to reflect on the overlap between this question and our earlier discussion of the problem of partial measures. That overlap comes from the fact that strategic adaptation can create both phenomena.

Remember what we are concerned about in the case of partial measures. Suppose you have only a partial measure of your mission (like hijackings). You take an action and things appear to improve on that measure. But there might have been strategic adaptation such that getting better on that one dimension of your mission implied getting worse on some other dimension. Hence, improvements on your partial measure may not mean improvements on your overall mission.

Strategic adaptation is again at the root when we worry about whether things will change once you act on some piece of evidence. There is some relationship in the world. You act on that relationship. People adapt to your actions. And, thus, the relationship disappears.

Many examples can fit into both categories, depending on how you think about them. Let us give you one last example to illustrate the point, this time about America's so-called war on drugs.

As we write, most of the illegal drugs in the United States enter the country through Mexico, a country that has been ravaged by a decade-long drug war. But this was not

always the case. In the 1970s and early 1980s, very few drugs reached the United States through Mexico. The transshipment route of choice was through the Caribbean and into Florida.

In 1980, the United States government launched a major offensive against the Colombian drug cartels. The Drug Enforcement Administration, Coast Guard, and other agencies deployed thousands of personnel and considerable naval and air power to shut down the Caribbean transshipment route. By the mid-1980s, the flow of drugs into Florida had plummeted.

But that is not the whole story. The reduction in drugs flowing through the Caribbean and into Florida in the 1980s does not reflect a reduction in drugs flowing into the United States during that period. Indeed, drugs continued to enter the United States at increasing rates, as evidenced by the fact that the price of cocaine plummeted fourfold during the course of the 1980s despite soaring demand.

So what happened? The Colombian cartels abandoned the Caribbean and Florida in favor of Mexico. In 1989, one-third of all cocaine in the United States entered through Mexico. Just three years later, that number had increased to one-half. Today, 90 percent of cocaine sold in the United States is smuggled up from Mexico.

This adaptation by the drug organizations has had devastating effects on Mexico. Throughout the 1990s, the Mexican drug trafficking organizations became larger and more powerful. They shifted from being middlemen for the Colombians to having their own suppliers and distribution networks. The drug trade became larger and more important—by the mid-1990s the Mexican drug trade was worth roughly $20 billion, dwarfing Mexico's largest legal commodity, oil, with a value of about $7.5 billion. As this expansion occurred, Mexican drug organizations became more fragmented and more violent. In 2010, the Mexican drug war claimed more than one thousand lives per month. The Mexican government struggled to exert basic control over parts of the country.

One can fruitfully think about this story from both the *partial measures* and the *changing relationship* perspectives.

From the *partial measures* perspective, here's how you'd tell it. The US government had a mission of stopping the drug flow. It noticed that almost all the drugs came up through the Caribbean. So it collected data—drugs flowing through the Caribbean—that was only a partial measure of the overall counter-narcotics mission. Then it took actions that made things improve according to that partial measure. But to conclude that the policy was a success would be a mistake. Because of strategic adaptation, getting better on that partial measure (drugs coming through the Caribbean) goes along with getting worse on other dimensions of the problem (drugs coming through Mexico). Here, the story illustrates the importance of not over-interpreting improvements on a partial measure of your mission.

From the *changing relationship* perspective, we tell the story slightly differently. There was a real correlation in the world—drugs were much more likely to enter the United States through the Caribbean than anywhere else. The government decided to act on the basis of this relationship by targeting its interdiction efforts in the Caribbean and Florida. Drug traffickers strategically adapted their behavior in response to this action, moving transshipment to Mexico. And so, as a consequence of the government's own actions, the correlation that formed the basis of the government's actions ceased to exist.

Both of these perspective are right. Which is more useful depends on the particular question you are trying to answer and the context in which you are trying to answer it.

Wrapping Up

Measuring your mission, like all the other lessons we've discussed, is an essential part of thinking clearly about how to use quantitative information to make better decisions. But, no matter how clearly you think, there are limits to what data and evidence can tell you. In chapter 17 we conclude the book by exploring some of those limits.

Key Words

- **Internal validity:** An estimate is internally valid if it is a credible estimate of the estimand (e.g., the estimator is unbiased).
- **External validity:** An estimate is externally valid if there is good reason to think the relationship will hold in a context other than the one from which the data is drawn.
- **Strategic adaptation:** Changes in behavior that result from an attempt to avoid the effects of a change in someone else's behavior.
- **Selected sample:** A sample of data that wasn't drawn at random from the population of interest but rather was selected to be studied because it possessed some particular set of characteristics.

Exercises

16.1 People who have already contracted COVID-19 and recovered from it are less likely to contract the disease again because of the immunity they have developed. However, a 2020 study published in the *The Lancet* suggests that those rare individuals who do contract the disease twice appear to experience worse symptoms the second time around. Using the thinking principles from this chapter and the fact that, because of limited testing, not all cases of COVID-19 are detected, provide an account of why this phenomenon might occur even if there is no biological mechanism that makes a second case of COVID-19 worse than a first case. Should this make you skeptical of the claim that people tend to experience worse symptoms when they contract the disease a second time?

16.2 Over the past several decades, high-stakes testing has become an increasingly important part of American education policy. The idea of high-stakes testing is to create consequences for students, teachers, or schools associated with performance on standardized tests. The hope is that this will improve educational achievement by creating incentives for better performance. Standardized test scores are, at best, a partial measure of educational achievement. Give an example of why some policy intervention that leads to an improvement in test scores might nonetheless not be leading to an overall improvement in educational outcomes.

16.3 In a required math sequence at the United States Air Force Academy, students take the same exams but are randomly assigned to different sections taught by different instructors. Scott Carrell and James West show that students assigned to an instructor with better teaching evaluations perform better on the course's exams. But they also show that being assigned to a popular instructor *decreases*

students' scores in subsequent math classes. What might explain this puzzling pattern? How does it relate to the problem of failing to measure your mission?

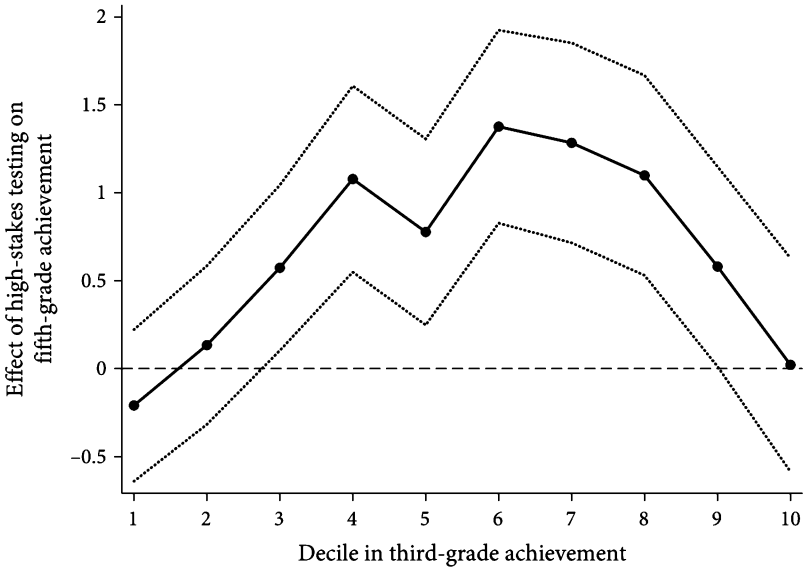
16.4 The way high-stakes testing is implemented in primary and secondary education is typically based on thresholds. A student passes the test if they get a score above some minimal cutoff. And a school is deemed to be meeting standards if the number of students passing the test is above some other minimal threshold.

(a) Think of students in three categories: those who will pass the test no matter what, those who will pass the test if and only if they get teacher attention, and those who will not pass the test no matter what. Which students does high-stakes testing of this form incentivize teachers to focus on?

(b) Derek Neal and Diane Whitmore Schanzenbach studied the implementation of high-stakes testing in the Chicago public schools that we discussed in exercise 1 of chapter 13.

But, unlike in that question, the average effect of high-stakes testing isn't quite what they wanted to know about. They wanted to know whether high-stakes testing affected different kids differently.

To get at this, Neal and Schanzenbach used a difference-in-difference-in-differences design. They started by using the third-grade tests to divide students into ten groups (deciles). They then perform a difference-in-differences analysis separately for each of these deciles. This allows them to learn about the difference in the difference-in-differences estimate of the causal effect of high-stakes testing across kids in the different deciles. Their findings are reflected in the figure.



Is this evidence consistent with your answer to part (a) of this exercise? Explain.

(c) In light of this, is a simple difference-in-differences design that uses the percentage of students passing the standardized tests a good way to evaluate whether high-stakes testing achieves its mission? Why or why not?

Readings and References

To read more about hijackings and metal detectors, see

Walter Enders and Todd Sandler. 1993. "The Effectiveness of Anti-Terrorism Policies: A Vector-Autoregression-Intervention Analaysis." *American Political Science Review* 87(4):829–44.

You can learn more about intermediate outcomes in medical research in

Thomas Fleming. 1994. "Surrogate Markets in AIDS and Cancer Trials." *Statistics in Medicine* 13:1423–35.

Thomas R. Fleming and David L. DeMets. 1996. "Surrogate End Points in Clinical Trials: Are We Being Misled?" *Annals of Internal Medicine* 125:605–13.

The study of temperature fluctuations and economic growth is

Marshall Burke, Solomon M. Hsiang, and Edward Miguel. 2015. "Global Non-Linear Effect of Temperature on Economic Production." *Nature* 527(7577):235–39.

For a comparison of the Tamil Nadu and Bangladesh Integrated Nutrition Projects, see

Howard White and Edoardo Masset. 2007. "Assessing Interventions to Improve Child Nutrition: A Theory-Based Impact Evaluation of the Bangladesh Integrated Nutrition Project." *Journal of International Development* 19(5):627–52.

The historical Major League Baseball statistics come from the Baseball Databank at seanlahman.com. The high school baseball statistics are from GameChanger at gc.com.

If you are interested in the history of adaptation to the window tax in England, have a look at

Andrew E. Glantz. 2008. "A Tax on Light and Air: Impact of the Window Duty on Tax Administration and Architecture, 1696–1851." *Penn History Review* 15(2).

Wallace E. Oates and Robert M. Schwab. 2015. "The Window Tax: A Case Study in Excess Burden." *Journal of Economic Perspectives* 29(1):163–80.

On the history of the shift in baseball, you can read

Travis Sawchik. 2017. "We've Reached Peak Shift." FanGraphs. November 9. blogs.fangraphs.com/weve-reached-peak-shift/.

The statistics on drug flows to the United States come from two reports:

United Nations Office on Drugs and Crime. 2010. "The Globalization of Crime: A Transnational Organized Crime Threat Assessment." Chapter 4. https://www.unodc.org/documents/lpo-brazil//noticias/2010/06/TOCTA_Report_2010_low_res.pdf.

Office of National Drug Control Policy. October 2001. "The Price of Illicit Drugs: 1981 through the Second Quarter of 2000." https://obamawhitehouse.archives.gov/sites/default/files/ondcp/policy-and-research/bullet_5.pdf.

The study of math exams and instructor quality in the Air Force Academy discussed in exercise 3 is

Scott E. Carrell and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118(3):409–32.

The discussion of high-stakes testing and teaching to students on the bubble in exercise 4 is most directly based on

Derek Neal and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92(2):263–83.

It also draws on

Bengt Holmstrom and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*. 7:24–52.