INTRODUCTION TO PSYCHOMETRICS

The two most common questions we get about our research are why we use surveys in our research (a question we will address in detail in the next chapter) and if we are sure we can trust data collected with surveys (as opposed to data that is systemgenerated). This is often fueled by doubts about the quality of our underlying data—and therefore, the trustworthiness of our results.

Skepticism about good data is valid, so let's start here: How much can you trust data that comes from a survey? Much of this concern comes from the types of surveys that many of us are exposed to: push polls (also known as propaganda surveys), quick surveys, and surveys written by those without proper research training.

Push polls are those with a clear and obvious agenda—their questions are difficult to answer honestly unless you already agree with the "researcher's" point of view. Examples are often seen in politics. For example, President Trump released his Mainstream Media Accountability Survey in February 2017, and the public

quickly reacted with concern. Just a few highlights from the survey underscore concerns with the questions and their ability to gather data in a clear, unbiased way:

1. "Do you believe that the mainstream media has reported unfairly on our movement?" This was the first question in the survey and is subtle, but it sets the tone for the rest of the survey. By using the term "our movement," it invites the survey respondent into an *us vs. them* stance. "Mainstream media" is also a negatively charged term in this political cycle.
2. "Were you aware that a poll was released revealing that a majority of Americans actually supported President Trump's temporary restriction executive order?" This question is a clear example of push polling, where the question tries to give the survey respondent information rather than ask for their opinion or their perceptions about what is happening. The question also uses a psychological tactic, suggesting that "a majority of Americans" support the temporary restraining order, appealing to the reader's desire to belong to the group.
3. "Do you agree with President Trump's media strategy to cut through the media's noise and deliver our message straight to the people?" This question includes strong, polarizing language, characterizing all media as "noise"—a negative connotation in this political climate.

We can see in this example why people could be so skeptical of surveys. If this is your only exposure to them, of course they can't

be trusted! No data from any of these questions can reliably tell what a person's perceptions or opinions are.

Even without an obvious example like push polling, bad surveys are found all over. Most often, they are the result of well-intentioned but untrained survey writers, hoping to gain some insight into their customers' or employees' opinions. Common weaknesses are:

- **Leading questions.** Survey questions should let the respondent answer without biasing them in a direction. For example, "How would you describe Napoleon's height?" is better than "Was Napoleon short?"
- **Loaded questions.** Questions should not force respondents into an answer that isn't true for them. For example, "Where did you take your certification exam?" doesn't allow for the possibility that they didn't take a certification exam.
- **Multiple questions in one.** Questions should only ask one thing. For example, "Are you notified of failures by your customers and the NOC?" doesn't tell you which part of the question your respondent was answering for. Customers? the NOC? Both? Or if no, neither?
- **Unclear language.** Survey questions should use language that your respondents are familiar with, and should clarify and provide examples when necessary.

A potential weakness of many survey questions used in business is that only a single question is used to collect data. Sometimes called "quick surveys," they are used quite often in

marketing and business research. These can be useful if they are based on well-written and carefully understood questions. However, it is important that only narrow conclusions are drawn from these types of surveys. An example of a good quick survey is the Net Promoter Score (NPS). It has been carefully developed and studied, is well-understood, and its use and applicability are well-documented. Although better statistical measures of user and employee satisfaction exist, for example ones that use more questions (e.g., East et al. 2008), a single measure is often easier to get from your audience. Additionally, a benefit of NPS is that it has become an industry standard and is therefore easy to compare across teams and companies.

TRUSTING DATA WITH LATENT CONSTRUCTS

With all of these things to watch out for, how can we trust the data reported in survey measures? How can we be sure that someone lying on their survey won't skew the results? Our research uses latent constructs and statistical analyses to report good data— or at least provide a reasonable assurance that data is telling us what we think it's telling us.

A latent construct is a way of measuring something that can't be measured directly. We can ask for the temperature of a room or the response time of a website—these things we can measure directly.

A good example of something that can't be measured directly is organizational culture. We can't take a team's or an

organization's organizational culture "temperature"—we need to measure culture by measuring its component parts (called manifest variables), and we measure these component parts through survey questions. That is, when you describe organizational culture of a team to someone, you probably include a handful of characteristics. Those characteristics are the component parts of organizational culture. We would measure each (as manifest variables), and together they would represent a team's organizational culture (the latent construct). And using survey questions to capture this data is appropriate, since culture is the lived experiences of those working on a team.

When working with latent constructs—or anything we want to measure in research—it is important to start with a clear definition and understanding of what it is we want to measure. In this case, we need to decide what we mean by "organizational culture." As we discuss in Chapter 3, the organizational culture that interested us was one that optimized trust and information flow. We referenced the typology proposed by Dr. Ron Westrum (2004), shown in Table 13.1.

Table 13.1 Westrum's Typology of Organizational Culture

Pathological (Power-Oriented)	Bureaucratic (Rule-Oriented)	Generative (Performance-Oriented)
Low cooperation	Modest cooperation	High cooperation
Messengers "shot"	Messengers neglected	Messengers trained
Responsibilities shirked	Narrow responsibilities	Risks are shared
Bridging discouraged	Bridging tolerated	Bridging encouraged
Failure leads to	Failure leads to justice	Failure leads to enquiry

scapegoating		
Novelty crushed	Novelty leads to problems	Novelty implemented

Once we have the construct identified, we write the survey questions. Clearly, the concept of organizational culture proposed by Dr. Westrum can't be captured in just a single question; organizational culture is a multifaceted idea. Asking someone "How is your organizational culture?" runs the risk of the question being understood differently by different people. By using latent constructs, we can ask one question for each aspect of the underlying idea. If we define the construct and write the items well, it works, conceptually, like a Venn diagram, with each survey question capturing a related aspect of the underlying concept.

After collecting the data, we can use statistical methods to verify that the measures do, in fact, reflect the core underlying concept. Once this is done, we can combine these measures to come up with a single number. In this example, the combination of the survey questions for each aspect of organizational culture becomes our measure for the concept. By averaging our scores on each item, we get an "organizational culture temperature" of sorts.

The benefit of latent constructs is that by using several measures (called manifest variables—the pieces of the latent variable that can be measured) to capture the underlying concept, you help shield yourself against bad measures and bad actors. How? This works in several ways, which are applicable to using system data to measure your system performance as well:

1. Latent constructs help us think carefully about what we

want to measure and how we define our constructs.
2. They give us several views into the behavior and performance of the system we are observing, helping us eliminate rogue data.
3. They make it more difficult for a single bad data source (whether through misunderstanding or a bad actor) to skew our results.

LATENT CONSTRUCTS HELP US THINK CAREFULLY ABOUT WHAT WE'RE MEASURING

The first way that latent constructs help us avoid bad data is by helping us think carefully about what we want to measure and how we are defining our constructs. Taking time to think through this process can help us avoid bad measurements. Take a step back and think about *what* it is you are trying to measure and how you will measure, or proxy, it. Let's revisit our example of measuring culture.

We often hear that culture is important in technology transformations, so we want to measure it. Should we simply ask our employees and peers, "Is your culture good?" or "Do you like your team's culture?" And if they answered yes (or no), what would that even mean? What, exactly, would that tell us?

In the first question, what do we mean by culture, and how did the respondent interpret it? Which culture are we talking about: Your team's culture or your organization's culture? If we really are talking about a workplace culture, what aspects of this work culture are we referring to? Or are we really more interested in

your national identity and culture? Assuming everyone understood the *culture* half of the question, what is *good?* Does good mean trusting? Fun? Or something else entirely? Is it even possible for a culture to be entirely good or entirely bad?

The second question is a bit better because we do specify that we're asking about culture at the team level. However, we still don't give the reader any idea of what we mean by "culture," so we can get data reflecting very different ideas of what *team culture* is. Another concern here is that we ask if the person *likes* their team culture. What does it mean to *like* a culture?

This may seem like an extreme example, but we see people make such mistakes all the time (although not you, dear reader). By taking a step back to think carefully about what you want to measure and by really defining what we mean by *culture*, we can get better data. When we hear that culture is important in technology transformations, we refer to a culture that has high trust, fosters information flow, builds bridges across teams, encourages novelty, and shares risks. With this definition of team and organizational culture in mind, we can see why the typology presented by Dr. Westrum was such a good fit for our research.

LATENT CONSTRUCTS GIVE US SEVERAL VIEWS INTO OUR DATA

The second way latent constructs help us avoid bad data is by giving us several views into the behavior and performance of the system we are observing. This lets us identify any rogue measures that would otherwise go undetected if they were the only measure

we had to capture the behavior of the system.

Let's revisit the case of measuring organizational culture. To begin measuring this construct, we first proposed several aspects of organizational culture based on Dr. Westrum's definition. From these aspects, we wrote several items.[1] We will talk about writing good survey items and checking them for quality in more detail later in the chapter.

Once we collect the data, we can run several statistical tests to make sure that those items do, in fact, all measure the same underlying concept—the latent construct. These tests check for:

- **Discriminant validity:** tests to make sure that items that are not supposed to be related are actually unrelated (e.g., make sure that items that we believe are not capturing organizational culture are not, in fact, related to organizational culture).
- **Convergent validity:** tests to make sure that items that are supposed to be related are actually related (e.g., if items are supposed to measure organizational culture, then they do measure organizational culture).

In addition to validity tests, reliability tests are conducted for our measures. This provides assurance that the items are read and interpreted similarly by those who take the survey. This is also referred to as internal consistency.

Taken together, validity and reliability statistical tests confirm our measures. They come before any analysis.

In the case of Westrum organizational culture, we have seven items that capture a team's organizational culture:

On my team . . .

- Information is actively sought.
- Messengers are not punished when they deliver news of failures or other bad news.
- Responsibilities are shared.
- Cross-functional collaboration is encouraged and rewarded.
- Failure causes inquiry.
- New ideas are welcomed.
- Failures are treated primarily as opportunities to improve the system.

Using a scale from "1 = Strongly disagree" to "7 = Strongly agree," teams can quickly and easily measure their organizational culture.

These items have been tested and found to be statistically valid and reliable. That is, they measure the things they are intended to measure, and people generally read and interpret them consistently. You'll also notice that we asked these items for a *team* and not for an organization. We made this decision when creating the survey items—as a departure from Westrum's original frame-work—because organizations can be very large and can have pockets of different organizational cultures. In addition, people can answer more accurately for their team than for their organization. This helps us collect better measures.

LATENT CONSTRUCTS HELP SAFEGUARD AGAINST ROGUE DATA

This deserves a slight clarification. Latent constructs *that are periodically retested with statistics and exhibit good psychometric properties* help us safeguard against rogue data.

What? Let us explain.

In the previous section, we talked about validity and reliability — statistical tests we can do to make sure the survey items that measure a latent construct belong together. When our constructs pass all of these statistical tests, we say they "exhibit good psychometric properties." It's a good idea to reassess these periodically to make sure nothing has changed, especially if you suspect a change in the system or environment.

In the organizational culture example, all of the items are good measures of the construct. Here is another example of a construct where tests highlighted opportunities to improve our measure. In this case, we were interested in examining failure notification. The items were:

- We are primarily notified of failures by reports from customers.
- We are primarily notified of failures by the NOC.
- We get failure alerts from logging and monitoring systems.
- We monitor system health based on threshold warnings (ex. CPU exceeds 90%).
- We monitor system health based on rate-of-change warnings (ex. CPU usage has increased by 25% over the last 10 minutes).

In preliminary survey design, we pilot-tested the construct with about 20 technical professionals and the items loaded

together (that is, they measured the same underlying construct). However, when we completed our final, larger data collection, we did tests to confirm the construct. In these final tests, we found that these items actually measured two different things. That is, when we ran our statistical tests, they did not confirm a single construct, but instead revealed two constructs. The first two items measure one construct, which appears to capture "notifications that come from outside of automated processes":

- We are primarily notified of failures by reports from customers.
- We are primarily notified of failures by the NOC.

The second set of items capture another construct —"notifications that come from systems" or "proactive failure notification":

- We get failure alerts from logging and monitoring systems.
- We monitor system health based on threshold warnings (ex. CPU exceeds 90%).
- We monitor system health based on rate-of-change warnings (ex. CPU usage has increased by 25% over the last 10 minutes).

If we had only asked our survey respondents if they monitor for failures with a single survey question, we would not be aware of the importance of capturing *where* these notifications come from. Furthermore, if one of these notification sources alters its behavior, our statistical tests will catch it and alert us. The same

concept can apply to system data. We can use multiple measures from our systems to capture system behavior, and these measures can pass our validity checks. However, we should continue to do periodic checks on these measures because they can change.

Our research found that this second construct, proactive failure notification, is a technical capability that is predictive of software delivery performance.

HOW LATENT CONSTRUCTS CAN BE USED FOR SYSTEM DATA

Some of these ideas about latent constructs extend to system data as well: They help us avoid bad data by using several measures to look for similar patterns of behavior, and they help us think through what it is we are really trying to proxy. For example, let's say we want to measure system performance. We can simply collect response time of some aspect of the system. To look for similar patterns in the data, we can collect several pieces of data from our system that can help us understand its response time. To think about what we are truly trying to measure—performance—we can consider various aspects of performance, and how else it might be reflected in system metrics. We might realize that we are interested in a conceptual measure of system performance which is difficult to measure directly and is better captured through several related measures.

There is one important note to make here: all measures are proxies. That is, they represent an idea to us, even if we don't acknowledge it consciously. This is just as true of system data as it

is of survey data. For example, we may use response time as a proxy for performance of our system.

If only one of the data points is used as the barometer and that one data point is bad—or goes bad—we won't know it. For example, a change to source code that collects metrics can affect one measure; if only that single measure is collected, the likelihood of us catching the change is low. However, if we collect several metrics, this change in behavior has a better chance of being detected. Latent constructs give us one mechanism to protect ourselves against bad measures or bad agents. This is true in both surveys and system data.

[1] These are commonly referred to as survey questions. However, they aren't actually questions; instead, they are statements. We will refer to them as survey items in this book.