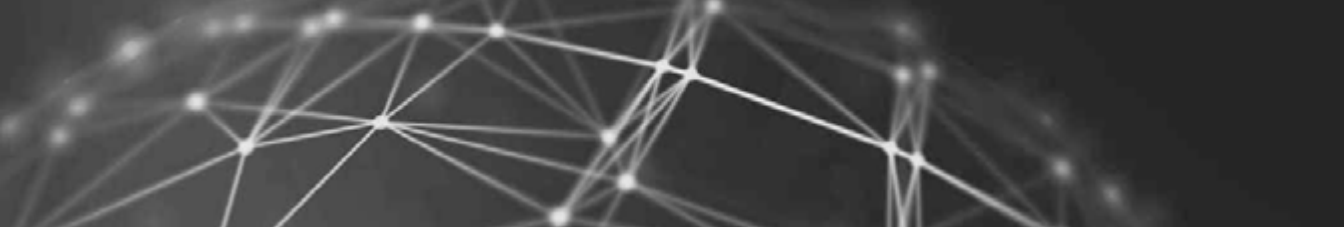

CHAPTER 2

Why Are You Measuring?

Connecting the Dots!

We have all sat through meetings where a variety of tables, charts, and graphs were presented for entirely more minutes than we thought we could ever endure. The presenter drones on in a monotonic voice describing the results of this month compared to last month or this quarter compared to the same quarter a year ago. You are bored but don't have the nerve to get up and walk out. So you endure. But you hold on to a vague hope that some minor crisis will erupt that will cause someone to call or send you a page with an urgent request for you to leave and attend to the pending crisis. Unfortunately, the call never comes. So, you endure the meeting till the end. As you are leaving the meeting you turn to a colleague and ask, "Did you get anything out of that meeting?" She looks at you and mumbles something that you believe sounds vaguely like "No, but that was certainly typical of how we attempt to connect the dots at our monthly management meetings." As you walk back to your office you realize that you and your colleague have just survived another journey into the "Data Zone"!

Okay, your Monday morning management meeting might not be quite like this one but it does bear a resemblance to many meetings that occur on a regular basis throughout healthcare settings. A frequent expression that arises in such meetings is, "You know what we need to do here? We need to connect the dots!" This is a rather popular expression that started me thinking. Why not give people a bunch of dots (data points) and ask them to predict what the image will be before we connect the dots? **FIGURE 2-1A** provides a series of dots. What do you predict will appear when the dots are connected? I have had many creative responses in class. The most frequent response is that the image will reveal a person riding a motorcycle. Others have seen a bear, a leaping animal, and even a rough image of their city, county, or country. Because no one has been able to accurately predict what the dots will produce when they are connected, I start to give them hints by connecting a few dots at a time. I show an image with dots 1–25 connected. I ask, "Now what do you see?" After a few more interesting guesses someone usually shouts out "an arm." Although this is correct, it does not give much insight into predicting what the remaining 158 dots will produce when connected. By the time I reveal the connections for the next 25 dots people are realizing that the image they are trying to predict is not a bear. At this point, however,

FIGURE 2-1A Connecting the dots (What image do you predict will emerge?)

they have narrowed down the range of possible outcomes but they still cannot predict what the dots will yield when connected. As I continue to connect more of the dots the class quickly comes to realize what the image is (see End Notes of this chapter for the answer if you have not figured it out by now).[1] They have now connected the dots and have produced an image of what the data points were hiding. So, the next image I show the class is Figure 2-1 again where we have all 183 dots but no lines. Of course they all see the image produced by the dots even though none of the dots are connected by lines. I then point out that once the image has been implanted in their brains it is hard not to see the final pattern created by the connected dots.

This simple example illustrates precisely what happens when people enter the "Data Zone." Data are presented, and the individual making the presentation explains what the dots are showing. In doing so, he or she is connecting the dots for you and in this process is imputing meaning to the dots that then sticks in your brain. You walk away thinking, "I guess the outpatient satisfaction scores are getting better." Or, "Wow I didn't realize that there was an upward trend in the inpatient fall rate." These impressions, however, may or may not be correct. The central point is that the dots have been connected for you by someone else so that when you walk out of the meeting you leave with an image and a conclusion. How many times have you been in a meeting where someone connected the data dots for you? When this happens do you raise your hand and ask whether there is another way to look at the data? Do you ask whether the presenter is merely describing what has happened in the past as opposed to predicting what we might expect the process to produce in the future? Do you politely ask whether there are alternative ways to analyze the data instead of showing two bar graphs and calculating a percentage change from Time 1 compared to Time 2? In short, can you predict what the dots are trying to tell you before they are connected? This is a key question for anyone interested in quality improvement (QI).

Data should be used to help you predict where you will be going in the future. Customers are not concerned about the average wait time last month or last quarter. They want to know why they are not being served now. A mother waiting in the emergency area, for example, wants to know when her febrile child will be seen by a doctor. She does not care one iota about the average wait time in the emergency area last week, last month, or last year. Similarly, a physician waiting for a STAT (immediate) troponin result takes no comfort in being told that the average turnaround time (TAT) last week was statistically significant from the average TAT last month and that the standard deviation went from 10.3 minutes to 9.4 minutes. In this case, the physician would have every right to voice concern.

The level of quality for a particular product or service is determined by understanding current performance and predicting the future not by how the process or system performed in the past. If you need an everyday example of how this works all you have to do is ask sports fans about the current performance of their favorite team, especially if that team won a championship in the past. I live in Chicago. We have a variety of professional sports teams including football,

basketball, baseball, ice hockey, and soccer. Several of these teams have won a championship within their respective sport. In 1985, the Chicago Bears professional football team won their one and only Super Bowl championship. Even today fans remember and reflect on the Super Bowl victory and talk with nostalgia about where they were and with whom when the Bears won. But that moment in 1985 does not help us predict the performance of the Bears this season, which was the worst in the history of the franchise. In 2016, a major sporting miracle occurred here in Chicago. Yes, after 108 years of failed seasons the Chicago Cubs won the World Series. The city went wild with excitement. But now that the celebration has ebbed some are already starting to talk about a repeat of this year's performance. Although hope springs eternal, people who make these types of statements have little understanding of a basic principle of human performance: past success is no guarantee of future success. Just ask the Bears or the Bulls fans!

Frequently, statistical methods reinforce this focus on past performance. The average (or any other descriptive statistic for that matter) for last month or last quarter provides an aggregated summary statistic that merely describes a single characteristic of past performance.[2] Dr. Myron Tribus, retired director of the Center for Advanced Engineering Study at Massachusetts Institute of Technology and a student of Dr. W. Edwards Deming, had a classic line about the problem with making comparisons based on monthly numbers: "Managing a company by means of the monthly report is like trying to drive a car by watching the yellow line in the rear-view window" (Wheeler, 1993, p. 4). Historical data by month, quarter, or year can be useful in helping you understand where you have been historically but it provides no basis for determining where you are right now (i.e., your baseline) or, more important, how you will perform in the future. Consider pilots and how they analyze data. You are flying in a Boeing 777 over the Atlantic from Chicago to London. You are having dinner and watching a movie on the tiny screen in the back of the seat in front of you. You are completely unaware of how much data the pilots are processing and the hundreds of dials and gauges they need to constantly monitor. They are processing data in real time to make predictions about the future. They are not sitting up front in the flight deck having this conversation: "Gee, Bill, what was our average fuel consumption per hour when we flew this same route last month?" "I don't know about fuel consumption, Tom, but I do have the average altitude and the standard deviation from projected altitude by quarter for the last 2 years." Because the pilots are concerned about where they are right now, the current flying conditions, and their ability to predict how the rest of the flight will go, the aggregated summary statistics from the past have little or no value.[3] From an improvement perspective data are meant to predict the future not describe past performance. The future can be understood only by analyzing repeated measures of the variables of interest over time and under a wide range of conditions.

Data and subsequently the interpretation of data require inquiry and dialogue. For example, when you are sitting in a meeting that feels like the one described at the beginning of this chapter you should realize that even though you are bored with the lengthy presentation of figures and statistics, the presenter is actually connecting the dots for you. When the presenter says, "And you can see from this summary table of numbers that the average wait time for registration, the average wait time to see the doctor, and the standard deviations for each of these measures have all dropped significantly over the past two quarters," your bored brain is actually connecting the dots. You walk out of the meeting with an image of what the presenter chose to tell you about the dots. You have no ability to predict what next quarter's wait times will be but you did make the connection that all the measures "dropped significantly over the past two quarters." The presenter used data that are aggregated and dated to plant a little causal model in your head. The dots have been connected. Lower wait times are better than high wait times and this past quarter was lower

than the previous quarter, therefore, we have improved. Right? The correct answer is that you do not know whether things have improved or not. All you have is two dots. If I would have connected only the first two dots in Figure 2-1 no one would ever have figured out the image. What about the other 181? Deming had a great line about two data points. In one of his 4-day seminars that I had the privilege of attending he said, "When you have two data points it is very likely that one will be different from the other."

Being able to predict where a process or a system will go in the future builds much more knowledge than merely describing what has happened in the past. In order to be successful at connecting the dots and using the data to make predictions about future performance, however, two questions need to be addressed:

- What type of study have you setup and what action(s) will result from the study?
- Why are you measuring performance?

These two questions are addressed in the remainder of this chapter.

Types of Studies

In 1938, Dr. Walter Shewhart delivered four lectures to the Graduate School of the Department of Agriculture in Washington, DC, on statistical thinking and quality control (QC). In 1939, he expanded on the content of these lectures and wrote *Statistical Method from the Viewpoint of Quality Control*. Shewhart outlined three components of knowledge (Shewhart 1939, 85–86):

1. The data of experience in which the process of knowing begins
2. The prediction in terms of data that one would expect to get if he were to perform certain experiments in the future.
3. The degree of belief in the prediction based on the original data or some summary thereof as evidence.

For Shewhart prediction was the central concept driving quality.

Dr. W. Edwards Deming (1942, 1950) built on Shewhart's initial thinking about prediction. In the foreword to *Quality Improvement Through Planned Experimentation*, Deming wrote this about prediction:

> Why does anyone make a comparison of two methods, two treatments, two processes, or two materials? Why does anyone carry out a test or an experiment? The answer is to predict—to predict whether one of the methods or materials tested will in the future, under a specified range of conditions, perform better than the other one. The question is, What do the data tells us? How do they help us predict? (Moen, Nolan, & Provost, 1991, p. xiii)

One's ability to predict, however, depends exclusively on the type of study designed and the actions to be taken as a result of the study. Deming provided a foundation for thinking about the types of studies that could be designed. He classified studies as being either enumerative or analytic (Deming, 1975). According to Deming (1975, p. 147), an enumerative study is one in which "action will be taken on the material in the frame being studied." He defined the frame as "an aggregate of identifiable tangible physical units of some kind, any or all of which may be selected and investigated" (p. 146). A classic example of an enumerative study is a census conducted on a country's population. In this case, the frame is the entire population of the country and the objective is to find out how many people live within the country's geographic boundaries. Following this definition, a frame in a healthcare setting might be all inpatients, all hip and knee replacement patients, a nursing unit, a long-term care facility, or all attending physicians.

Three key points highlight enumerative studies:

1. The aim of an enumerative study is principally descriptive in nature

(Deming, 1975, p. 147; Wheeler, 1995, p. 18). Enumerative studies basically describe how many or how much and are essentially based on historical data. The focus of an enumerative study, for example, is not on explaining why there were more males than females receiving a particular medication but rather on merely quantifying how many males and females received the particular medication. Prediction of the future is not possible with enumerative studies.

2. The actions to be taken or the decisions to be made as a result of an enumerative study will be (or should be) directed only to the subjects in the frame.[4]
3. Random sampling methods based on probability theory and related techniques such as confidence intervals and tests of significance usually provide the statistical approach to enumerative studies.

In contrast to the enumerative study is an analytic study. Deming defined an analytic study as one "in which action will be taken on the process or causal system that produced the frame studied with the aim being to improve practice in the future" (1975, p. 147). A key distinction between an enumerative and analytic study, therefore, is that analytic studies focus on the *future performance* of the system or frame being studied, in other words prediction. A fundamental assumption of analytic studies is that the conditions that produce an observable outcome today will be different tomorrow and the next day and the next. Enumerative studies, on the other hand, look at data that have occurred in a defined period of time (e.g., last month, last quarter, or even last year) that is fixed or static in nature.

Provost (2011) and Provost and Murray (2011) provide a useful analogy for thinking about this important distinction of the role of time in study design. The analogy is to water quality in a pond versus a moving stream or river. They compare an enumerative study to a pond that is fixed and not moving. One or more random samples drawn from the pond are sufficient to determine the water quality of the entire pond. They point out that traditional statistical methods such as hypothesis testing (i.e., rejecting or not rejecting the null hypothesis), confidence intervals, and tests of significance can be used to make decisions about the quality of the water in the pond based on random sampling techniques. But note that the conclusions and subsequently the decisions about the pond in question must be restricted to only that pond. If there is a pond of similar size 15 yards away from the tested pond, no conclusions should be made about that pond even though they are relatively close to each other.

The writers' analogy for an analytic study is a stream or river that is in constant motion. If you are standing along a rapidly moving stream, for example, the water quality and the properties in the water right in front of you are gone in a second and new water is now in front of you. Does the water now in front of you have the same properties and qualities as the water that just went past a few seconds ago? Thus water quality determined by a random sample of the water at a single point in time will not be able to reveal anything about the quality of the water that will pass in front of you 10 seconds later. The stream is in constant motion. It is not fixed or static like the pond. The present condition of the water in the stream will change over time. For example, imagine that a chemical plant discharges its waste into the stream at 2 a.m. every other day. If you drew a single random sample of water from the stream at 2 p.m. on the day between when the chemical plant dumps its waste and took this sample back to the laboratory for testing you would most likely miss the chemical waste that was dumped at an entirely different point in time. As Wheeler (1995, p. 18) points out, "there is no way to define a random sample of the future." Without repeated sampling over time it is impossible to detect a change in the water quality. So, whereas enumerative studies rely heavily on random sampling methods, analytic studies use primarily judgment sampling techniques, both of which will be discussed in greater detail in Chapter 4.

Similar analogies can be found throughout the healthcare industry. Patients connected to telemetry in the intensive care unit (ICU), for example, are monitored in an analytic not an enumerative fashion. We connect ICU patients to a variety of electrical leads in the ICU that allow us to track their heartbeats, respiration rate, and blood pressure in real time and over time. Why do we do this? Why don't we just take one random sample of their blood pressure at some random point during their 11-day stay in the ICU and use this one reading to make clinical decisions about the patient's progress? It would save time for the nurses and most likely save the ICU money. Or maybe we could get their average of all blood pressure readings on the first day of admission to the ICU then hold off till they are discharged 10 days later and take the second average of all blood pressure readings on this last day. We could even go further and discover whether the two average blood pressure readings between admission and discharge are statistically different by applying a test of significance to determine whether the resulting difference between admission and discharge is significant at the 0.05 or 0.01 level of significance. If we took this enumerative approach to patient care we would end up causing considerable harm to the patients. Clinical decision making is essentially grounded in analytic study designs. We track patient conditions over time, especially when the need to know necessitates moment-to-moment monitoring as in the ICU. The medical model is clearly more analytic in nature than enumerative. Yet, the distinctions between these two approaches are typically not typically taught in medical, nursing, allied health professions, or healthcare administration programs.

One final distinction between enumerative and analytic studies needs to be reinforced before moving on. Specifically, it is critical to realize that each approach employs different statistical methods to arrive at conclusions. Enumerative studies rely on probability theory, descriptive and comparative statistics, and tests of significance. If you have ever had to take a basic or even advanced course in statistics, these are the methods and techniques you have been exposed to. In some circles these area referred to as "traditional statistics." These include both univariate (descriptive) as well as multivariate statistical methods, which includes everything from tabular analysis (i.e., crosstabs) to various forms of regression analysis, factor analysis, cluster analysis, discriminant analysis, and analysis of variance. A major reference from the social science field for these approaches can be found in the works of Hubert Blalock (1979) and in particular his classic textbook *Social Statistics*.

When conducting analytic studies graphical methods of analysis are used most often. This branch of applied statistics is generally referred to as statistical process control (SPC) and was developed by Dr. Walter Shewhart in the early 1920s (more about this in Chapter 9). SPC methods, most notably the run chart and the Shewhart control chart, accommodate repeated sampling over time that in turn allows the researcher to understand the variation inherent in the process and thus predict the future performance of the process. More advanced statistical methods of prediction are incorporated in analytic studies by using multifactorial designs and planned experimentation (Moen Nolan, & Provost, 2012). Because this book is concerned principally with analytic studies these methods will be discussed in detail in the remaining chapters.

Research for Efficacy, Efficiency, and Effectiveness

Closely aligned with the type of study you design is a very pragmatic question: Why are you measuring? This seems like a simple and straightforward question. Yet many healthcare professionals do not think about why they are actually measuring. You may hear managers or frontline workers, for example, say, "Look, we need to submit some data on our progress related to how much time patients have to wait

before actually seeing the doctor in the family practice clinic so find some recent wait time numbers and send them in." Frequently this means the data submitted may not be the most recent data, it may not be defined in the same way it was defined when it was first submitted last year, or it may not be presented in a format that adequately answers the questions being posed. Anyone engaged in performance measurement, therefore, needs to be very clear about their reasons for starting the quality measurement journey (QMJ).

Brook, Kamberg, and McGlynn (1996) provided guidance for helping healthcare professionals think about why they are measuring. They differentiated two basic research paths that can help individuals determine the purpose of their measurement efforts: research for efficacy and research for efficiency and effectiveness. The choice of which path is to be followed should be based on the nature of the questions the researcher is trying to answer.[5] All scientific inquiry begins with questions not with statistical tools.

The efficacy road takes the researcher down the traditional enumerative path of scientific inquiry described previously. This path is grounded in experimental and quasiexperimental designs (Campbell and Stanley, 1966; Posavac and Carey, 1980; Weiss, 1972). In health care, efficacy studies are frequently used to test the ability of a particular drug, treatment, procedure, or protocol to improve medical conditions. For example, a researcher might pose the following questions about a new drug:

- Is this drug capable of producing the desired effect?
- Will this drug act differently with different types of patients (age, gender, race, etc.)?
- Does a 5-mg dose of the drug produce different results than a 10-mg dose?
- What does the literature tell us about the use of this drug?
- What does past research reveal about the drug's application in experimental or quasiexperimental trials?

Every time healthcare researchers conduct a randomized clinical trial (RCT) to compare the impact of a new drug or a protocol, they are conducting research for efficacy purposes. A typical study might involve testing the effect of a new blood pressure drug on two groups of patients. One group would receive the drug and be labeled as the experimental group. The other group would receive a placebo and would be referred to as the control group. Baseline blood pressure readings would be obtained on each group, as well as demographic information about their current physiological condition, family history, and activities of daily living, etc. After a period of time, the two groups would have their blood pressure levels measured again, and statistical tests would be performed to see whether there is a "significant" difference between the two groups. The demographic and other patient characteristics would be used as control variables (i.e., they are used in an effort to "hold constant" confounding or exogenous variables that might play a role in blood pressure variation).

If we really wanted to increase the precision of this type of study, we would establish matched samples of patients (i.e., we would make every possible effort to have the members of the experimental and control groups be similar in gender, age, race, socioeconomic status, and so on). The participants would be randomly assigned to the experimental and control groups, and then steps would be taken to make the study a double-blinded trial. The double-blinded component means that neither the participants nor the researchers know which group gets the drug and which one gets the placebo. This is done in order to minimize bias, increase the validity of the results, and thereby enhance the researcher's ability to answer the efficacy question.

The study design I have just described is a classic approach to conducting a research study. In a design of this type, the goal is to test the null hypothesis (H_o) that there is no difference between the experimental and control groups with respect to the blood pressure drug. Statistical analysis of the before and after blood pressure readings for

the two groups allows the researchers to reject or not reject the null hypothesis and test for statistical significance (Babbie, 1979; Morrison & Henkel, 1970; Selltiz Jahoda, Deutsch, & Cook, 1959).

Although some traditional research studies are based on time series analysis or interrupted time-series analysis (McDowall McCleary, Meidinger, & Hay, 1980; Ostrom, 1978), the majority of the studies designed to test efficacy are not designed to monitor moment-to-moment fluctuations in an observed process or outcome. Instead, they typically obtain rather large sample sizes (e.g., 100 or more observations in two comparison groups), let weeks or months transpire as the research trial is allowed to run its course, and then see whether the two comparison groups show a significant statistical difference. The focus of most experimental and quasiexperimental designs, therefore, is on static comparisons (i.e., enumerative designs that make comparison of datasets that are fixed at discrete points in time).

In summary, the primary purpose of research for efficacy purposes is to build knowledge by testing theories against empirical evidence. The research results may not have immediate or even readily apparent practical outcomes, however. For example, research done in the vacuum of space by astronauts has produced amazingly pure crystals, but scientists are not sure what they can actually do with this new knowledge on earth. This new knowledge will be added to the rest of the body of knowledge about crystalline formation, and someday someone will figure out how to use this knowledge for practical purposes. In the meantime, articles will be written and new theories about the nature of crystals will be explored.

When healthcare professionals receive training in research designs and statistical methods, the approach described here is usually the standard frame of reference (i.e., research for efficacy). In my mind, this is one of the primary reasons why many healthcare professionals seem to have a strong proclivity to focus on comparing two numbers (e.g., this month compared to last month or this quarter compared to the previous quarter) and asking whether one time period is statistically different from the other. Research for efficacy purposes is designed to answer this type of question.

Brook et al. (1996) refer to the second research pathway as one that leads to insights about *efficiency* and *effectiveness*. This is the road that leads to QI research and is closely aligned with analytic study designs described earlier in this chapter. Research related to efficiency and effectiveness is also consistent with what Westfall, Mold, and Fagnan (2007) and Khoury et al. (2007) call translational research.

Closely linked with the ideas of Shewhart (1931), Deming (1992), and Juran (1988, 1992), research for efficiency and effectiveness takes a very practical approach that seeks to answer questions about the variation in process or outcome indicators over time instead of whether one number is different from another (i.e., the efficacy question). The key characteristics of research for efficiency and effectiveness include:

- A focus on solving practical problems rather than testing or building theoretical models.
- A dynamic perspective rather than a static or aggregated perspective. This is the distinction discussed earlier about the difference between enumerative and analytic studies.
- Use of small samples of data selected continuously over time. For example, sample sizes could be as small as five patients a day or 10 each week. The frequency of sampling when conducting research for efficiency and effectiveness depends on the unit of time (i.e., the subgroup) to be used in the study. More will be said about this point in Chapter 4.
- More reliance on graphical displays of data than on descriptive statistics and tests of significance.

Since Brook et al. made the distinction between research for efficacy and research for efficiency and effectiveness in 1996, considerable

progress has been made to incorporate the latter form of research into the healthcare field. Today, what Brook called research for efficiency and effectiveness is essentially known as the science of improvement (SOI). The rich history of the SOI has been summarized nicely by Moen and Norman (2010). Perla, Provost, and Parry (2012) provide additional depth in exploring what they call the "seven propositions of the SOI." Taylor et al. (2013) provide a very nice review of how the SOI has been applied to improvement efforts in healthcare settings.[6]

The key messages Brook et al. (1996) helped to reinforce were that all measurement is the same and that the scientific method lies at the heart of all good research, whether it is done on large static comparison groups (i.e., efficacy research or enumerative studies) or in real-time settings (i.e., research for efficiency and effectiveness or analytic studies). All research designs and methods have utility. It is up to the individual researcher to build a knowledge base that allows him or her to know which design, method, and/or statistical technique is most appropriate for the questions that need to be answered. Consider an analogy from the surgical field. The surgeon has many tools and methods available for surgery. The challenge is not to use all of them just because they are there or to state unequivocally that one tool or method is superior to all the others. Instead, the real challenge is to have knowledge of all the tools and methods and know when to use the right tool at the right time to solve the problem at hand. It is the same way with research. The question or problem being addressed should be the primary driver for deciding which research design or methods are appropriate for turning data into information. Some research methods and tools are best suited to answer questions related to efficacy. Others are more appropriate for questions related to effectiveness and efficiency. The wise (and in my opinion humble) researcher should know enough about each approach to know when to use one approach and not the other. The questions, not the statistical tools or methods, should drive the research endeavor.

The Three Faces of Performance Measurement

In 1997, Solberg, Mosser, and McDonald wrote an excellent article that provided a very practical compliment to the more conceptual framework provided by Brook et al. (1996). Solberg et al. identified what they called the three faces of performance measurement: measurement for improvement, measurement for accountability (or what I refer to as judgment), and measurement for research. They further argue that these three approaches to measurement should not be mixed: "We are increasingly realizing not only how critical measurement is to the QI we seek but also how counterproductive it can be to mix measurement for accountability or research with measurement for improvement" (p. 135).

The authors describe the characteristics of each of the three faces and how each approach to measurement is based on a different purpose or aim and methods. **TABLE 2-1** provides my adaptation of the key points from this valuable article. The three faces are listed as the column headings and the rows identify the major aspects that need to be addressed during the measurement journey. The reader is encouraged to spend a few minutes reviewing the details in each cell of this. Even a quick perusal of the table's content will quickly reveal that the three faces not only have very different aims but also use different methods and strategies to collect and analyze data. For each of the three faces the following key characteristics are briefly discussed:

- Measurement aim
- Testing methods and observability
- Data collection and sample size
- Determining whether the data demonstrate an improvement or change (I don't get why these words were capped and the previous lines were not? So I changed them to lower case.)

TABLE 2-1 The three faces of performance measurement

Aspect	Improvement	Accountability (Judgement)	Research
Aim	Improvement of care (efficiency and effectiveness)	Comparison, choice, reassurance, motivation for change	New knowledge (efficacy)
Methods: ■ Test observability	Test observable	No test, evaluate current performance	Test blinded or controlled
■ Bias	Accept consistent bias	Measure and adjust to reduce bias	Design to eliminate bias
■ Data	"Just enough" data, small sequential samples	Obtain 100% of available relevant data	"Just in case" data
■ Flexibility of hypothesis	Flexible hypothesis, changes as	No hypothesis	Fixed hypothesis (null hypothesis)
■ Testing strategy	Sequential tests	No tests	One large test
■ Determining whether change is an improvement	Analytic statistics (SPC); run & control charts	No change focus (maybe computer a percent age change or rank the order of results)	Enumerative statistics (t-test, F-test, chi square, p-values)
■ Confidentiality of the data	Data used only by those involved with improvement	Data available for public consumption and review	Research subjects' identity protected

Modified from Lief Solberg, Gordon Mosser and Sharon McDonald, Journal on Quality Improvement vol. 23, no. 3, (March 1997), 135-147. By permission of Robert Lloyd.

As you review Table 2-1 notice that I have placed dashed lines as the vertical dividers between the three columns. This was done to help remind us that the three approaches should not be separate and isolated from each other. The learning from each column should be permeable and influence the other two columns. More will be said about this notion later.

The improvement column is where many healthcare professionals claim to be working. Yet, I often find that although it is quite popular to say you are committed to quality and safety these days, many organizations do not follow the measurement principles and practices that Solberg et al. laid out a number of years ago. How well do your QI measurement activities align with the following characteristics?

- *Measurement Aim.* As mentioned previously, the aim of measurement for improvement

is centered on enhancing the efficiency and effectiveness of care processes and outcomes. There is a direct connection between measurement for improvement and measurement for research. Improvement is basically the extension of the research act. Traditional research helps us determine "the what" (i.e., what might be a reasonable new idea to test?) whereas improvement research allows us to determine "the how." How can we implement an efficacious idea, technique, procedure, or drug so that it performs 100% of the time in an efficient and effective manner every time it is administered or applied?

- *Testing Methods and Observability.* When a team is working on testing a new idea that they feel will improve a process or outcome they are usually following what Campbell and Stanley (1963, p. 37–43) would classify as quasiexperimental time series design. This is not as rigorous or complex a design as a true experimental design but it is quite good for addressing issues of internal validity, which is what QI is all about (i.e., comparing your current performance against a more desirable level of performance over time). The other characteristic of improvement measurement related to methods is that the work of the researchers (i.e., the team) is observable not only to the team and management but to anyone else who wants to see the team's work. It is not uncommon for example, to observe team members posting the results of their most recent tests in a location that the patients or the public can review.[7]
- *Data Collection and Sample Size.* Data collection for improvement is structured around small samples collected sequentially within time periods as close to the actual productions of work as possible. Solberg et al. describe this approach to data collection as "good enough" data collection. They write, "Because a high degree of precision is not necessary for improvement purposes and because data collection needs to be simple and repetitive, small samples, for example, 10–20 cases per sample are usually appropriate" (Solberg et al., 1996, p. 142). This approach to data collection is referred to as gathering "just enough" data.
- *Determining whether the data demonstrate an improvement or change.* (Again these capitals) the primary branch of statistics used to analyze improvement data is SPC. This branch of analytical statistics is a combination of graphical displays of data over time plus statistical estimates of the variation within the data display. This is achieved by analyzing the data with run charts or Shewhart charts. Note that statistical tests of significance (e.g., p-values) are not appropriate for improvement research. On this point Deming (1992, p. 312) wrote "Students are not warned in class nor in the books that for analytic purposes (such as to improve a process), distributions and calculations of mean, mode, standard deviation, chi-square, t-test, etc. serve no useful purpose for improvement of a process unless the data were produced in a state of statistical control. The first step in the examinations of data is accordingly to question the state of statistical control that produced the data." Details on these statistical methods are provided in Chapters 8 and 9.

The second column in Table 2-1 addresses measurement for accountability or as I refer to it, judgment. The role of measurement for judgment within the healthcare industry has increased dramatically over the past 10 years throughout the world (see Chapter 1 for additional detail on this topic). Measurement for accountability (judgment) should not always be labeled as negative. We all need to be accountable for our actions and the work we produce. It is important for both internal as well as external purposes. The problem from my perspective is when individuals are judged, rated, or ranked and such assessments are not linked to improvement. The key characteristics of measurement for accountability and judgment are:

- *Measurement Aim.* The primary purpose of measurement for accountability or judgment is to make comparisons, rate

and rank performance, pass judgment on performance, help customers make choices (as is done for example by *Consumer Reports* or the Leapfrog Group in the United States that rates and ranks healthcare providers), decide on the distribution of bonuses and incentives, and/or drive competition and stimulate a desire for change. This form of measurement is typically focused on comparing groups or individuals and asking a very simple question, "Is your performance better now than it was the last time we looked at you? Yes or no?" In most instances, the answer to this question is based on current performance compared to past performance or performance against targets or goals.

- *Testing Methods and Observability.* There is no rigorous testing occurring within this approach to performance measurement let alone any application of experimental or quasiexperimental designs. Because the focus is on evaluating current performance against past performance all that matters is whether there is a difference between time 1 and time 2. When the comparisons are made it is not uncommon to use national, regional, state or provincial norms as comparative reference markers (dare I say benchmarks?) for external comparisons and targets and goals for internal comparisons.
- *Data Collection and Sample Size.* The objective is to obtain 100% of available data for the defined period of time, which is usually by quarters or years and is usually aggregated into summary statistics. The major controversy that emerges from this approach is that the data are lagged and the aggregated results, therefore, may not reflect current performance. For example, hospital ratings and rankings at the state, province, or national level may lag for as much as a year or more. In these cases when the results are released hospital leaders are quick to point out that the current conditions of operation and their results are quite different than they were a year ago when the comparative data were collected. The final aspect of data collection for judgment is that most of the measures in this category are outcome measures (e.g., overall hospital mortality, infection rates, or patient responses to a general survey question such as, "How would you rate the overall quality of your care?"). The problem created by a singular focus on high-level outcome indicators is that a focus on outcomes alone provides little or no insight into the processes that produce the outcomes or more important what needs to be changed in order to move the outcomes to a more preferred level of performance. This occurs most often when external agencies or organizations pass judgment on healthcare providers. When accountability or judgment is done internally against target or goals, however, there is usually more concern over hitting internal targets or goals that can be aimed at improving the processes that drive the outcomes.
- *Determining Whether the Data Demonstrate an Improvement or Change.* Statistical analysis for accountability and judgment is essentially a binomial question. Are you better now than when we last looked at you? Yes or no? Do you have fewer infections now than when we posted your data a year ago? Has your inpatient mortality dropped? Yes or no? The answer to this question is usually based on comparing raw numbers, percentages, or rates at two points in time. Some comparisons will be based on calculating a percentage change to see whether it meets or exceeds a specified target or goal. One of the more popular methods used to determine whether a change has occurred from a judgment perspective is to use what is popularly known as the "traffic light" scorecard. This approach establishes three cut points against a target or goal and then

assigns red, yellow, or green colors to the units of observation based on where each unit falls against the targeted cut points. In this case, green is usually at or above target, yellow indicates that the unit of observation is not getting better or worse, and red indicates performance below target or goal. A variation on this same methodology is to assign stars rather than colors to indicate the level of performance. The better the performance the more stars you get. Such methods to determine whether a change has occurred are frequently tied with pay for performance bonus and incentive programs. Although the negative impacts of these approaches have been well documented by Deming (1992, 1994), Kohn (1986, 1993), Berwick (1995), and Herzberg (2003), the traffic light method of analyzing performance is still widespread in healthcare settings. Besides being antithetical to the basic tenets of QI, the problem with the traffic light approach is that it completely ignores the underlying variation in the processes and related outcomes being judged. Further details on understanding variation will be explored in Chapter 6.

The third face of performance measurement is measurement for research purposes. This approach serves a vital and extremely important function within the healthcare industry. Interestingly enough, however, although healthcare professionals frequently reference RCTs and research findings there are actually very few healthcare professionals who work full time in this area.[8]

- *Measurement Aim.* As has been stated the primary aim of research (or efficacy to use both Solberg's and Brook's term), is to develop new theories, test existing theories, and build knowledge.
- *Testing Methods and Observability.* We take elaborate steps to help ensure that exogenous variables that could confound or mess up our study and thereby invalidate the results are controlledor held constant. Experimental and quasiexperimental designs (Campbell & Stanley, 1963) are used to maximize the validity and reliability of the results and reduce threats to the study.[9] Frequently, blinded tests are used so that neither the researchers nor the participants in the study know which participants receive the actual test intervention and which ones receive the placebo.
- *Data Collection and Sample Size.* One of the key points of difference between the three faces of performance measurement relates to data collection strategy. Solberg et al. point out that when conducting research studies we typically create large, complex datasets that are based on data that have occurred in the past (i.e., last quarter or more often last year). The authors refer to this type of data as "just in case" data. I believe this term is used because when doing research we usually collect more data than we may need "just in case" reviewers of our research, journal editors, or critics raise questions about our results or methods. If this happens, we can respond, "No problem. I collected additional data 'just in case' you raised that question."[10]
- *Determining Whether the Data Demonstrate an Improvement or Change.* When we conduct statistical analysis for research we typically use methods that were mentioned previously in the description of enumerative studies. Usually these approaches compare groups or identify which variables are believed to have "significant" influence over the dependent or outcome variable. In a research context we usually employ statistical tests of significance and most notably a p-value to determine whether the null hypothesis (fixed hypothesis) is rejected or not and significant results are observed between the experimental and control groups. Although the statistical notion of "significance" has played a major role in making clinical decisions, there has been considerable debate in the literature about the "significance test controversy" (Ioannidis, 2005;

FIGURE 2-2 Viewing the three faces of performance measurement as silos and as a Rubik's cube

Zwarenstein & Oxman, 2006; Morrison & Henkel, 1970; Ziliak & McCloskey, 2011). Readers interested in this topic should make a special effort to read the works of Ziliak and McCloskey. They provide a wonderful historical review of the development of the test of significance, especially the p-value, and then provide a comprehensive review of the debate surrounding the test of significance controversy. In their overview of the three faces of performance measurement Solberg et al. conclude "how counterproductive it is to mix measurement for accountability or research with measurement for improvement" (1997, p. 135). This may have been true in 1997 but I do not think it is quite true today or an appropriate approach for current time and challenges. In my opinion, healthcare organizations need leaders and frontline individuals who are comfortable and competent in being able to blend the three faces of performance measurement. Although I agree with the authors that it can send confusing messages to say you are doing research and then apply procedures and methods that characterize measurement for accountability (judgment) or improvement or any other combination of the three faces and the aspects of measurement, I do not view the three approaches as separate and distinct silos as shown in **FIGURE 2-2**. I also do not believe that it is "counterproductive to mix measurement for accountability or research with measurement for improvement." Instead I believe that it is more effective to view the three faces as a Rubik's cube. **FIGURE 2-3** provides a graphic image of this analogy.

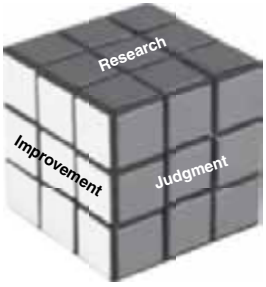

FIGURE 2-3 Viewing the three faces of performance measurement as a Rubik's cube

Healthcare leaders will improve their measurement capacity if they navigate the edges and interfaces of the three faces of performance measurement. Sometimes, a person becomes strongly invested in only one type of measurement. A person may talk about

measurement for research as the "only" valid type of measurement for example. Similarly, others may talk about aggregate and summary data as being preferable because it allows them to compare hospitals, clinics, and groups of doctors, cities, or regions. In this case, they are invested squarely in the center of the measurement for judgment space of the cube. Others may say that although the other approaches to measurement must be addressed, the preferable way to proceed with measurement is for QI and all other approaches are of secondary importance. Instead of creating more silos within the healthcare industry we should be investing time in creating what Konan (1981) calls translation. According to Konan (p. 12), "The term 'translation' is chosen because the process is like translating prose from one language to another. A good language translation is not a literal rewriting of each sentence. Rather it captures the essence of the original copy and then uses the mentality of the second language to present the material to a different audience." If we are to have translation between the three faces of performance measurement, then we need translators not silo builders. Konan picks up on this notion by stating, "Being a good foreign-language translator requires more than the ability to read and write another language. Similarity, the skills required for translating research into everyday language are different from those required for doing research. The translator must be able to read and understand scientific research and then synthesize and interpret it in relation to policy perspectives" (p. 13). Healthcare organizations need individuals who can function as translators and be able to talk to individuals who are leading each of the three faces of performance improvement and find the linkages between improvement, judgment, and research. It is not that difficult but it does require the development of specialized skills as Konan points out.

For physician and nurse leaders especially, the edges of the cube where the three faces are connected are critical. They need to be able to walk comfortably back and forth between the three faces of the cube so as to understand how valid and reliable measurement is structured and organized within each of the three approaches. Translators are not stuck in any one of the measurement silos but work to break down barriers and then build on the strengths of each approach. All three approaches must be understood as a system. The problem is that individuals identify with one of the approaches and dismiss the value of the other two. This leads to duplication of effort and redundancy in data collection and places limitations on learning. A short case study demonstrating how a translator with the appropriate skills and knowledge can flip the cube around to address the aims of each of the three faces while using the same data follows.

CASE STUDY #1: *Being a Translator*

Situation

The board of a 310-bed community hospital is concerned that the hospital did not fare too well in a recent citywide report on inpatient infection rates. This report compared 16 hospitals and showed that the hospital ranked 13th out of the 16 hospitals. The report was recently highlighted in a newspaper story and the hospital was singled out as one of the worst performers. An improvement team was established to improve these results.

(continues)

CASE STUDY #1: *Being a Translator* *(continued)*

The Data

The QI department worked with the 11 inpatient units and 2 critical care units to establish historical baselines on the number of infections using run charts with the number of infections (e.g., central line bloodstream infections, clostridium difficile, and urinary tract infections) plotted on the vertical axis and week displayed as the unit of time on the horizontal axis. Each unit then initiated various improvement strategies and continued to track the various infections on their run charts. Each time they tested a new idea they annotated the run chart to identify when the test was conducted.

Discussion

As the teams tracked their improvement tests and plotted the weekly number of infections, they noticed that a majority of the units demonstrated a gradual decline in the number of infections. The three units that did not show declines in the number of infections visited the other units to determine what they were doing that did produce better results. After several team meetings these two units also started to notice declines in their numbers of infections. So now it was time to report back to management and the board. The staff in the QI department, functioning as translators, knew that the board had a very simple question for them: "Do we have fewer infections now than when we first saw the ranking of the hospital in the public report?"

The QI staff had all the detailed improvement data for the 13 units plotted by week on the run charts. They knew, however, that this was too much detail for the board. So, they flipped the Rubik's cube around to the accountability face and aggregated the detailed run charts into two numbers; the median number of infections for the entire hospital in the baseline period and the median number of infections over the last 12 weeks. The baseline median for the entire hospital was 29 infections whereas the more recent median after changing the infection assessment protocol and management processes was 17. Both the baseline and the current number of infections were below the number that had been printed in the public report. This allowed the QI director to go to the board meeting and provide a simple answer to the board's question: "Yes, we are safer now than we were when the public report was released and we have continued to improve since then." The QI translators have been able to take the detailed improvement data from the units, flip it around, aggregate it, and provide the board with an answer to their accountability question.

The final step in translating these data would come about when the translators take the detailed improvement data and the run charts from the units and meet with researchers to discuss and possibly design a quasiexperimental research study to see whether a totally new intervention that had not been tested previously could be conducted to determine whether the idea is efficacious. In this way, the same dataset has been used to address questions relevant to all three faces of performance measurement. The key, however, is having translators who are able to understand the aims and methods of each of the three faces and then be able to develop linkages between the sides of the Rubik's cube.

We have covered a lot of ground in this chapter. We started by connecting the dots and thinking about our ability to predict with data. This led us into a discussion on the types of studies with a focus on the differences between enumerative and analytic studies. Next we thought about the roles of research for efficacy as compared with research for efficiency and effectiveness. This led us finally to discussing the three faces of performance measurement and the fundamental question "Why are you measuring?" With these fundamental concepts in place we are now ready to start our own QMJ.

Notes

1. Figure 2-1B with all 183 dots connected. Did you predict that it would be a hurdler?

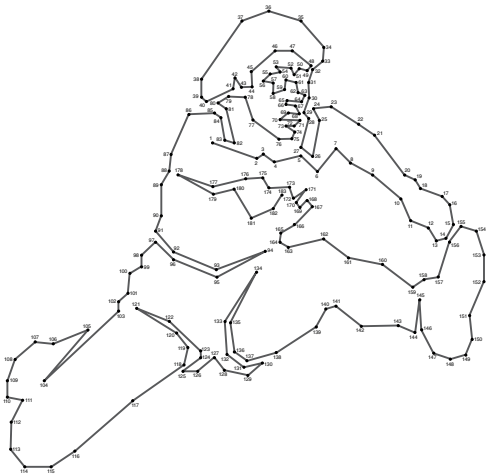

FIGURE 2-1B All 183 dots connected. Did you predict that it would be a hurdler?

2. You will remember from your basic statistics class (or "sadistics" as some might call them) that there are two basic groups of descriptive statistics. First, there are measures of central tendency characterized by the mean, the median, and the mode. The mean is simply the arithmetic average of all the numbers in a distribution. The median is the point where half the data are above this point and half are below it. Another name for the median is the 50th percentile. The mode is the most frequently appearing number in a distribution of numbers. In the classic picture of the theoretical normal curve these three numbers are shown as the vertical line in the center of the bell curve and depict the mean, median, and mode all at the same position. Typically this happens only when a theoretical normal curve is being shown or when computer simulations are run. In real life, however, it is rare for the mean, median, and mode to all be represented by the same number. The second group of descriptive statistics consists of measures that reveal the dispersion in a distribution of data. The simplest of these measures is the maximum and minimum values in a distribution. This is followed by the range, which is the absolute difference between minimum and maximum values. Finally there are a number of descriptive statistics that capture various aspects of dispersion. These include the sum of the deviations, the average deviation, the sample or population variance, and the standard deviation. Further details on any of these descriptive statistics can be found in any basic statistics text.
3. This is not to say that the historical data have no value at all. They have little or no value to the pilots making the flight you are on right now. But the historical data are of great value to the designers, engineers, and technicians at Boeing who built and maintain the 777s. The historical data on past performance is also of value if there is an accident and the Federal Aviation Administration is trying to determine when a possible failure started to occur. The key point to be made in this example is that the question you are trying to answer should guide your measurement journey. The measures you choose to track, your data collection plan, statistical analysis techniques, and the conclusions you can make all start with the question(s) you are trying to answer.
4. This is a sometimes a challenge for individuals who conduct enumerative studies. Specifically they do not confine themelves to making inferences about the members or entities in the frame. As one of the professors in my doctoral program used to say, "They frequently want to drive beyond their headlights." That is they define a frame and then arrive at a conclusion about the members or entities in the frame but they don't stop there as they should. If, for example, an enumerative study is conducted on all the hip replacement surgeries done

last year at Hospital A and the researchers report that 62% of the patients who participated in pre-hab (i.e., engaging in rehabilitation exercises prior to surgery to strengthen their muscles) were able to walk with the assistance of only a cane on the second day after surgery then all they can rightfully conclude is that this is the outcome for Hospital A. If there are two other hospitals in the system, however, that also do hip replacement surgery but were not part of the original frame would the researchers be justified in telling the other hospital orthopedic directors that they should have all their hip replacement patients engage in pre-hab because Hospital A had 62% of its patients walking with a cane on the second day after surgery? If they did arrive at this conclusion they would be "driving beyond their headlights." Their study merely described a characteristic of a subset of patients in the frame. This conclusion provides no insights as to why 62% of the patients were walking with assistance of a cane on the second day. Who were these patients? All male? A mix of males and females? What ages were these patients? Did they have any comorbidities? When questions like these are being asked, the study needs to move from being an enumerative to an analytic design.

5. Note that I am using the term "researcher" in a fairly broad context in this text. For me, you do not need an academic degree or work for an organization that has the word research in its name to be called a researcher. I maintain an applied perspective when it comes to defining who might be a researcher. Anyone who is interested in answering a question, finding a better way to perform a task or activity, or generating new knowledge for themselves or others is in my mind a researcher.
6. Despite the rich history of the SOI and its very direct links to the scientific method, it is interesting that the SOI is frequently challenged as not being a "legitimate science." When I hear this type of statement I always wonder if the person making the comment would ask sociologists, psychologists, economists, or chemists to prove that their disciplines are scientific in nature.
7. Observability or transparency of improvement tests and the related results are not always widely embraced, especially in a country as litigious as the United States. Some healthcare organizations in this country, for example, have been counseled by their legal staff to not reveal, report, or publicly post QI results. The concern is that if a patient experienced an adverse event or harm and data are posted publicly then the family and their lawyers might use the data to support their claim that the provider did not have quality processes in place at the time of the incident. This position has been shown to have little credibility, however, and the transparency of QI initiatives is increasing with fewer legal challenges each year (see, e.g., Baily et al., 2006).
8. I have arrived at this conclusion by asking participants in my workshops how many of them spend a majority of their time working on setting up and running experimental or quasiexperimental medical research projects. Even in audiences of several hundred I usually have either only a few or no hands go up. Most healthcare professionals do not spend a majority of their time in conducting academic research studies designed for efficacy purposes. On the other hand, when I ask audiences how many of them spend most of their time being accountable for results or trying to improve results nearly all the participants have their hands up in the air. The point is that although we rely on research to move the healthcare field forward, a majority of healthcare professionals, those delivering care on a day-to-day basis, are not engaged in conducting research. These people are

being charged with being accountable for and/or improving the efficiency and effectiveness of care.

9. One of the best resources on this point is Campbell and Stanley (1963). This is a classic reference relevant to any field of research and something that in my opinion should be required reading for any healthcare professional. The authors discuss three preexperimental designs, three true experimental designs, three quasiexperimental designs, counterbalanced designs, and four separate-sample pretest–posttest designs. They discuss each of these designs in terms of 12 threats to internal and external validity, which are a challenge to be addressed in any study design. Some of the designs handle these threats better than others. For example the weakest of all designs is the one-shot case study. Campbell and Stanley point out that this design is used frequently in educational studies. I would add to this that it is also used unfortunately all too often in healthcare settings. You will probably recognize this study design. It carried out with a single study group over two time periods. Time period 1 is the pretest marker, sometimes referred to as a baseline measure. The intervention is administered to the study group and then a posttest measurement of the variable of interest is taken. For example, imagine that your long-term care facility has been ranked in the bottom decile of a statewide study on falls in nursing homes. The board of your facility has made it clear to the management team that this outcome is unacceptable and must be improved. You have the pretest result showing that your facility is in the bottom decile. Someone comes up with the idea of placing posters up throughout the facility admonishing everyone to "Pay Attention! Only **YOU** can prevent patient falls!" The posters are in place for 1 week. Then the number of falls is recorded each day for the next week and the average is calculated. This serves as the posttest value. Although your pretest number of falls over the past year was on the average 23 per month, there is great excitement when the posttest average (which remember is only for 1 week) reveals only 16 falls. What do you conclude from these two numbers? The wrong conclusion is that the posters "caused" fewer falls. The correct conclusion is that it reveals absolutely nothing. Campbell and Stanley write that "such studies (i.e., one-shot case studies) have such a total absence of control as to be of almost no scientific value" (p. 6). Yet we see many pretest–intervention–posttest designs in health care in which completely incorrect conclusions are offered. Most healthcare study designs fall into the category of preexperimental designs. This is why I believe that all healthcare professionals, especially managers and leaders, need to spend a little time learning from Drs. Campbell and Stanley.

10. An example of this "just in case" approach to data collection comes from my own experience. When I was building the dataset for my doctoral dissertation I spent the better part of a year designing, organizing, and executing the data collection plan. My dissertation research was designed to identify factors that led manufacturing firms to locate in rural and small communities in Pennsylvania (Lloyd, 1983). My data plan involved collecting data from Pennsylvania state archives, the U.S. Census, Dun and Bradstreet's Market Identifiers File, County Business Patterns, the Pennsylvania Bureau of Economic Analysis, local government documents, and phone call flow data from AT&T. This was clearly a "just in case" dataset that enabled me to respond not only to additional analytic questions my committee posed to me but also to address editorial questions when it came time to publish the results of my research in a professional journal (Lloyd & Wilkinson, 1985).

References

Babbie, E. R. *The Practice of Social Research.* Belmont, CA: Wadsworth Publishing Company, 1979.

Baily, M., M. Bottrell, J. Lynn, and B. Jennings. "The Ethics of Using QI Methods to Improve Quality and Safety in Health Care." *Hastings Center Report* 36, no. 4 (2006): S1–S40.

Berwick, D. "The Toxicity of Pay for Performance." *Quality Management in Health Care* 4, no. 1 (1995): 27–33.

Blalock, H. *Social Statistics*, rev. 2nd ed. New York: McGraw-Hill, 1979.

Brook, R., C. Kamberg, and E. McGlynn. "Health System Reform and Quality." *JAMA* 276, no. 6 (1996): 476–480.

Campbell, D. and J. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin, 1966.

Deming, W. E. "On Classification of the Problems of Statistical Inference." *Journal of the American Statistical Association* 37 (1942): 173–185.

Deming, W. E. *Some Theory of Sampling.* New York: Wiley & Sons, 1950, reprinted in 1960.

Deming, W. E. "On Probability as a Basis for Action." *American Statistician* 29, no. 4 (1975): 146–152.

Deming, W. E. *Out of the Crisis.* Cambridge, MA: Massachusetts Institute of Technology, 1992.

Deming, W. E. *The New Economics*. Cambridge, MA: Massachusetts Institute of Technology, 1994.

Herzberg, F. "One More Time: How Do You Motivate Employees?" *Harvard Business Review* 81, no. 1 (2003): 86–96.

Ioannidis, J. "Why Most Published Research Findings Are False." *PLoS Medicine* 2, no. 8 (2005): 124. doi:10.1371/journal.pmed.0020124

Juran, J. *Juran on Planning for Quality*. New York: Free Press, 1988.

Juran, J. *Juan on Quality by Design.* New York: Free Press, 1992.

Khoury. M. J., M. Gwinn, P. W. Yoon, N. Dowling, C. A. Moore, and L. Bradley. "The Continuum of Translation Research in Genomic Medicine: How Can We Accelerate the Appropriate Integration of Human Genome Discoveries into Health Care and Disease Prevention?" *Genetics in Medicine* 9, no. 10 (2007): 665–674.

Kohn, A. *No Contest: The Case Against Competition*. New York: Houghton Mifflin Company, 1986.

Kohn, A. *Punished by Rewards.* New York: Houghton Mifflin Company, 1993.

Konan, M. "Translation: A Neglected Stage on the Research Process." *Rural Sociologist* 1, no. 1 (1981): 11–18.

Lloyd, R. "Vertical and Horizontal Linkages on Manufacturing Employment In Rural Communities of Pennsylvania." PhD diss., The Pennsylvania State University, 1983.

Lloyd, R., and K. Wilkinson. "Community Factors in Rural Manufacturing Development." *Rural Sociology* 50, no. 1 (1985): 27–37.

McDowall, D., R. McCleary, E. Meidinger, and R. Hay. *Interrupted Time Series Analysis.* Beverly Hills, CA: Sage Publications, 1980.

Moen, R., T. Nolan, and L. Provost. *Quality Improvement Through Planned Experimentation*, 3rd ed. New York, McGraw Hill, 2012.

Moen, R., and C. Norman. "Circling Back: Clearing up Myths about the Deming Cycle and Seeing How It Keeps Evolving." *Quality Progress*, November 2010, 22–28.

Morrison, D., and R. Henkel, eds. *The Significance Test Controversy: A Reader.* Chicago: Aldine, 1970.

Ostrom, C. *Time Series Analysis: Regression Techniques*. Beverly Hills, CA: Sage Publications, 1978.

Perla, R., L. Provost, and G. Parry. "Seven Propositions of the Science of Improvement: Exploring Foundations." *Quality Management in Health Care* 22, no. 3 (2012): 170–186.

Posavac, E., and R. Carey. *Program Evaluation: Methods and Case Studies*, 4th ed. Englewood Cliffs, NJ: Prentice Hall, 1980.

Provost, L. "Analytic Studies: A Framework for Quality Improvement Design and Analysis." *BMJ Quality & Safety* 20, supplement 1 (2011): i92–i96.

Provost, L., and S. Murray. *The Health Care Data Guide*. San Francisco, CA: Jossey-Bass, 2011.

Selltiz, C., M. Jahoda, M. Deutsch, and S. Cook. *Research Methods in Social Relations.* New York: Holt, Rinehart & Winston, 1959.

Shewhart, W. *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand, 1931; reprinted by the American Society for Quality, 1980.

Shewhart, W. *Statistical Method from the Viewpoint of Quality Control.* Mineola, NY: Dover Publications, 1939.

Solberg, L., G. Mosser, and S. McDonald. "The Three Faces of Performance Measurement: Improvement, Accountability and Research." *Journal on Quality Improvement* 23, no. 3 (1997): 135–147.

Taylor, M., C. McNicholas, C. Nicolay, A. Darzi, D. Bell, and J. E. Reed. "Systematic Review of the Application of the Plan–Do–Study–Act Method to Improve Quality in Healthcare," *BMJ Quality & Safety* 23, no. 4 (2013): 1–9. doi:10.1136/bmjqs-2013-001862.

Weiss, C. *Evaluation Research: Methods of Assessing Program Effectiveness*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Westfall, J., J. Mold, and L. Fagnan. "Practice-Based Research—Blue Highways on the NIH Roadmap." *JAMA* 297, no. 4 (2007): 403–406.

Wheeler, D. *Understanding Variation: The Key to Managing Chaos*. Knoxville, TN: SPC Press, 1993.

Wheeler, D. *Advanced Topics in Statistical Process Control*. Knoxville, TN: SPC Press, 1995.

Ziliak, S., and D. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor, MI: University of Michigan Press, 2011.

Zwarenstein, M., and A. Oxman. "Why Are So Few Randomized Trials Useful, and What Can We Do About It?" *Journal of Clinical Epidemiology* 59 (2006): 1125–1126.