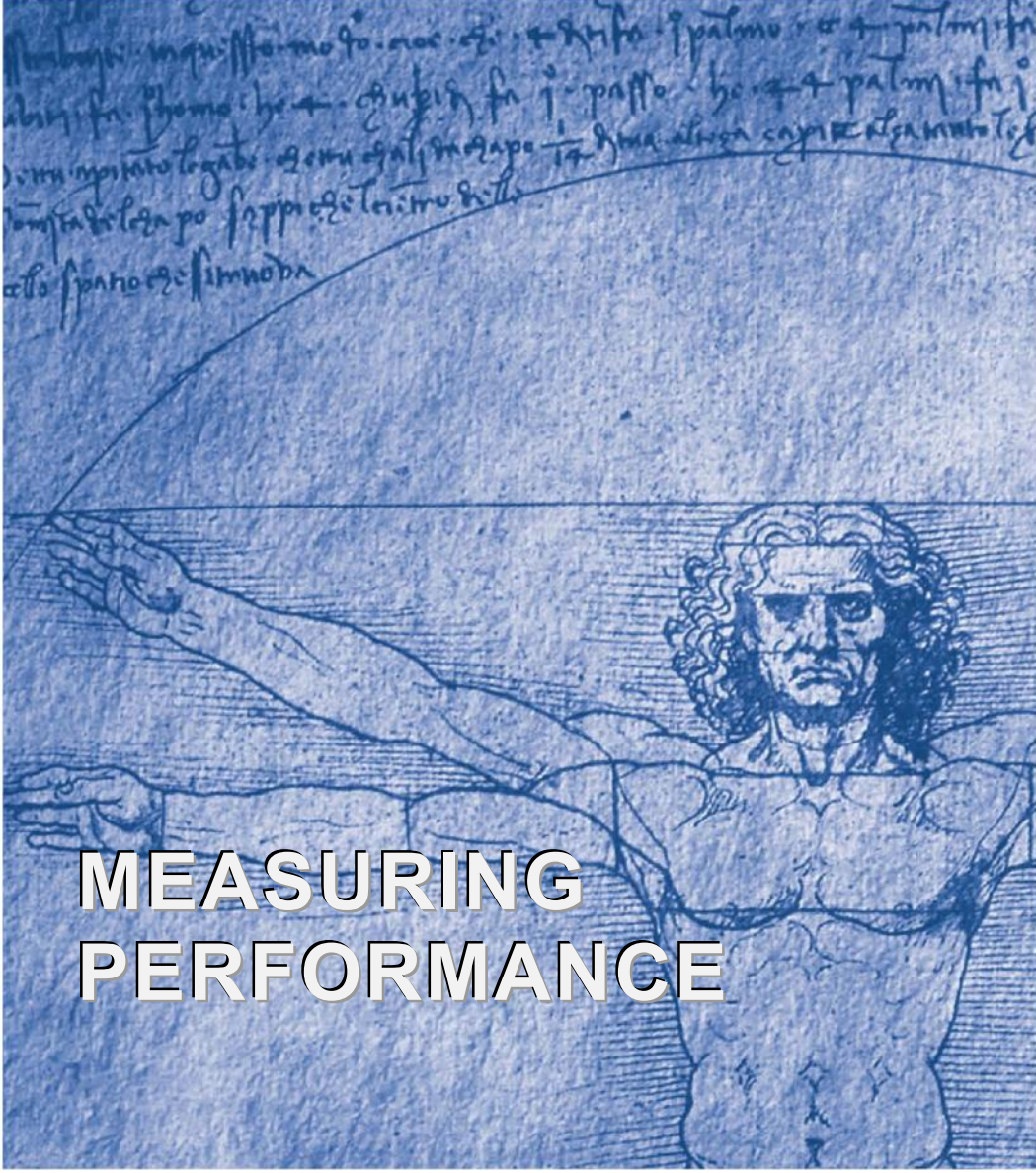


MEASURING PERFORMANCE

Damned if you do and damned if you don't

ASQ
Human Development & Leadership Division

This primer is brought to you by ASQ's Human Development & Leadership Division. Our mission is to *'to be the community of choice for everyone by making human potential a global priority, an organization and personal imperative.'*

We serve the community by providing publications like this, education, webinars, conferences and other resources for personal and professional growth and for leadership skills development. To learn more about us, you can visit us at:

http://asq.org/hdl/about/awards-hdl.html

Please contact me at adil@pinnacleprocess.com if you need more information or are interested in working with our team.

Adil F. Dalal, Chair, HD&L Division, ASQ

MEASURING PERFORMANCE:

Damned if you do and damned if you don't

Brooks Carder, PhD

I have always seen the plan function in the plan-do-study-act cycle as a place where data is important. While it is not always possible to obtain relevant data, it is more often possible than most decision makers suspect. Of course data can come from a formal scientific study, or from the informal questioning of a few key people. By key, I mean key to the process under study, not your boss or some other high-ranking official, unless they have information about the process. In fact the collection and analysis of relevant data is critical to successful process improvement. This primer is focused on performance measurement, because that is a universal challenge. But the principles apply to any measurement process.

The title of this primer is intended to serve two purposes: An obvious one is the marketing angle, being a bit outrageous, and hopefully creating attention and interest. The other one, perhaps less obvious, is to warn the reader of the potential perils of performance measurement. I am reminded of the Heisenberg uncertainty principle. This principle applies to quantum physics, not to employee performance reviews. One simple way of describing the principle is that when you measure the momentum of a particle, you lose the ability to know where it is going, because the measurement process changes the trajectory of the particle. But likewise, when you measure the performance of a system there is a substantial likelihood that you will change the trajectory of that system. And as we shall see that change is not always for the better.

A brilliant consultant once said," What gets measured is what gets done." To some extent that is correct, but I would add that “what gets measured is often the wrong thing, so that what gets done is often

the wrong thing." Of course, this applies principally to measures that are used to evaluate the performance of individuals and groups.

But why do we measure performance anyway? Perhaps we have an innate urge to do so. A popular item on Yahoo's homepage is the top 10 of something. It might be the top 10 cities for new jobs, the top 10 cities for affordable housing, the 10 most negatively viewed celebrities, or whatever. The point is people want to see who won or who is ahead.

Of course there are many other reasons to measure performance. We might want to record progress. We might want to make informed choices such as which ballplayer to draft or what company to invest in. We might even want to develop information that would assist us in a process improvement effort.

All of these appear to be legitimate applications of performance measurement. So what's the problem? Why are we damned if we measure performance? We will see later in the primer how performance measurement can give rise to cheating, manipulating the numbers, and focusing on the numbers rather than the process. Attempting to evaluate the performance of employees, in the form of annual performance reviews creates an additional problem, as these reviews can be very demoralizing to the employees. Such reviews are often a waste of time anyway, since the performance of employees tends to be quite dependent on the system in which they work. We shall see dramatic evidence of this. However there is no doubt that we will continue to measure the performance of organizations and individuals. The point is that we need to be constantly aware of these limitations. I shall elaborate on them later.

Six topics

This primer will focus on six topics:

1. How the quality of a measure is assessed through its reliability and validity.

2. How measuring performance often leads to manipulation of the numbers and/or cheating, and how to identify when this is happening.

3. How the performance of individuals is largely determined by the system in which they perform. Often attempting to measure individual performance is a waste of time, or worse.

4. How using reduced data (averages, medians, etc.) may lead you to miss important information.

5. How focusing on the wrong measure can lead to disaster.

6. How it is important to use more than one measure when measuring a complex process.

Assessing the quality of a measure

The quality of a measure is determined by its reliability and validity.

Reliability. A measure is reliable if repeated attempts to perform that measurement yield similar results. Realize they will not yield identical results, as all measures have variation. In fact a lack of variation is a sign that something is wrong with the measurement process. Deming often pointed out that "there is no true value of anything." He noted that the speed of light is an important constant in physics, but that the number assigned to it depends upon the method by which it was measured. Rather than there being a true value of anything there is a method for measurement and a result of

that method. If the results are relatively repeatable then we determine that the measure is reliable.

People often mistake the apparent concreteness of the measure with reliability. I have spent a considerable amount of time in the measurement of safety performance. Often we compare two methods: counting accidents, and surveying safety culture. An accident would appear to be a very concrete event that should be easy to identify. A survey of attitudes and beliefs is not as concrete. Most line managers trust accident counts and don't trust surveys. But it turns out that accident counts can be quite unreliable for a number of reasons.

Determining whether something should be counted as an accident is not as simple as it would seem. The criterion lines tend to get moved if the unit has already had too many accidents. In small organizations or units the control limits of accidents are so wide that the unit might have one accident every five years. In four of the years they would have an outstanding safety record and the fifth-year their record would be absolutely unsatisfactory. I am not inventing this. I've seen it happen. If management had used to control chart, of course, the process would've been shown to be in control even in the year the accident occurred. But many organizations do not plot accidents on control charts.

It turns out that our safety surveys are extremely reliable. There are two common methods to assess the reliability of a survey. One is the split-half method by which you randomly assign each question to one half or the other and then compare the scores of the two halves. A second method is to compare the scores of the unit in one year to the scores of that same unit in another year. With the survey, the split half test yields coefficients on the order of .9, which is excellent.

Figure 1 is a scatter plot of a number of plants tested in 1996 and 1997. There is a strong correlation, 0.82. Since the surveys were a year apart, we would not expect the correlation to be much higher, as the sites undergo changes in that period.

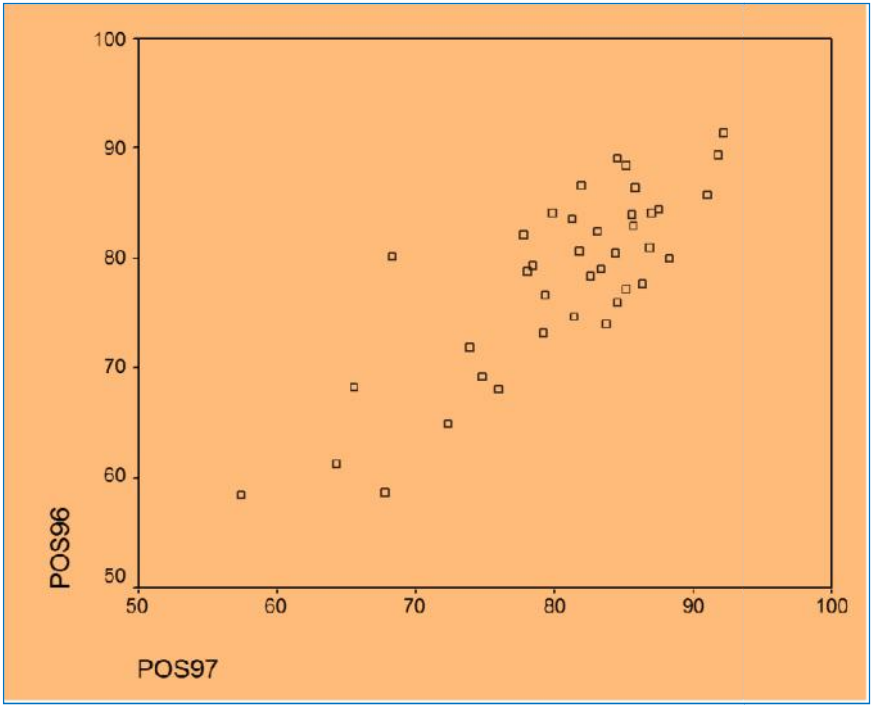


Figure 1

The point here is that you shouldn't assume that the measurement method is reliable because it seems simple and concrete. You need to actually measure the reliability.

Face validity. The first kind of validity is called face validity. If you want to measure safety, counting accidents is a logical step. It has face validity. If you want to measure the performance of a

corporation, looking at things like the stock price and its profitability are logical steps. They have face validity. Realize however, that just because a measure has face validity does not mean that it's a useful measure. There are two other kinds of validity that are probably more important. In fact you might discover a very useful measure that doesn't have much face validity. If it has the next two kinds of validity it would be useful.

Figure 2 is a section of the safety survey we are referring to. You can see that most of questions have clear face validity.

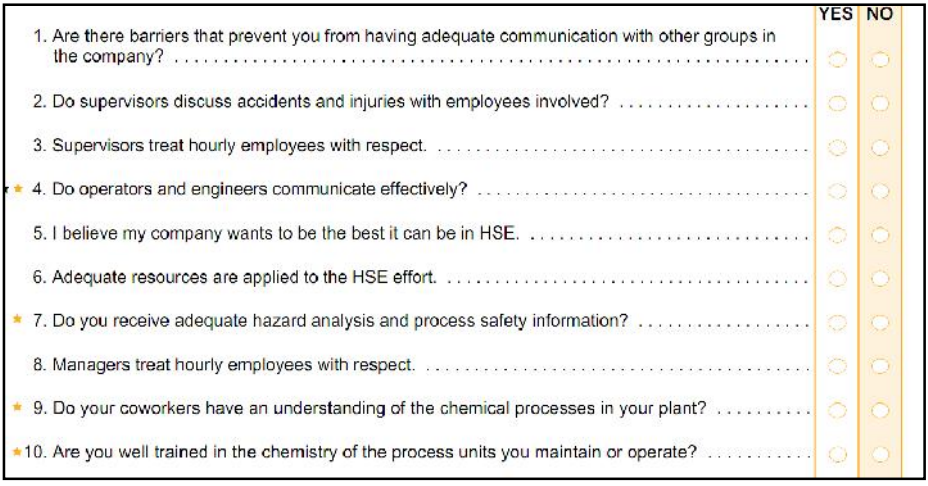


Figure 2

Predictive validity. An example of measurement that has good predictive validity but not so much face validity is found in baseball. In his book *Moneyball*, Michael Lewis describes how the Oakland Athletics were able to develop successful teams with a much smaller payroll than teams like the Yankees and Red Sox. Baseball insiders, for example, believed that the most important criterion for a pitching prospect is how fast he can throw the ball. Thus prospects who could

throw at high speeds were in high demand and consequently were very expensive. Baseball statisticians had discovered that, independent of throwing speed, a pitcher who could get batters out in college could also get batters out in professional baseball. Thus slow throwing but successful college pitchers were a bargain. Baseball insiders, such as the scouts for most teams, stuck to their belief in the face validity of throwing velocity. Baseball insiders, of course, are the arbiters of face value.

A measure with predictive validity correlates with other measures. For example, IQ test scores correlate with academic success and even with economic success. However, one has to be aware of the magnitude of that correlation. One form of IQ test is the Standardized Aptitude Test (SAT) test which is used by many colleges to determine who should be admitted. The reason for using this is that high school grades are not comparable from one high school to another. Yale University has used the SAT test for many years. When I spoke to the admissions office there several years ago they told me that the correlation between SAT scores and Yale grades was on the order of .2 to .3. This means that the SAT score accounts for less than 10% of the variation in grades at Yale. The other 90% is accounted for by things like motivation, work habits, the difficulty of courses taken, etc. While the SAT score is not a powerful predictor, it is the best they have.

I and my colleagues have had a great deal of experience with safety surveys. To measure the predictive validity of individual questions, we compared the scores from some excellent sites, as judged by low accident rates over three years, and expert evaluations, with the scores of weak sites (high accident rates and low evaluations.) Validation requires a statistically significant difference on a Chi-square test.

For validation of the whole survey we correlated the survey score against three year accident rates and against expert evaluations.

Figure 3 shows a scatter plot of 13 sites. Survey scores are plotted against accident rates.

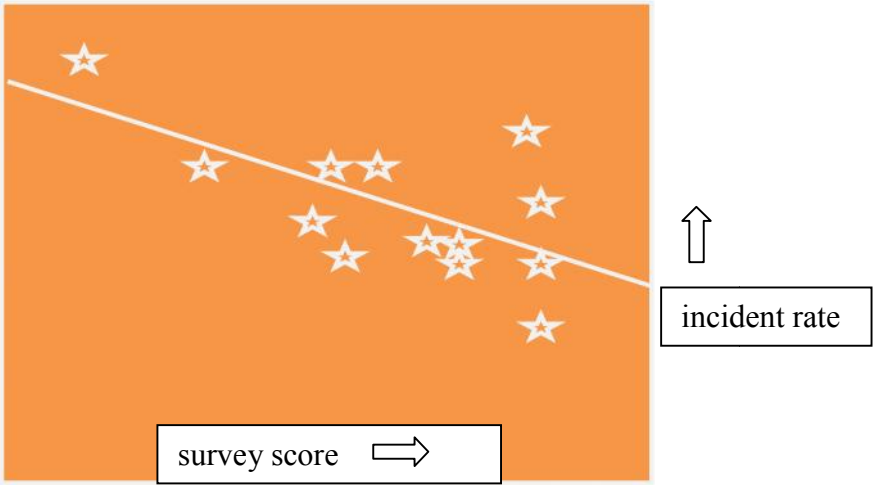


Figure 3

The correlation is as expected. In general the higher the survey score, the lower the incident rate. The correlation coefficient is 0.64 which is quite strong, and highly significant.

Construct validity. A measure with construct validity gives you information about the thing you are measuring beyond simply performance. A measure with good construct validity will help you to develop improvement plans. Consider the IQ test. It has predictive validity but lacks any construct validity. If a person has a low IQ score there is no prescription for improvement. Many performance measures have little construct validity. Ideally you want a measure that will assist you in developing an improvement plan. Our safety

surveys, for example, have strong construct validity. The survey comes with a defined process for subsequent action. This process involves feeding the results back to the workforce, assembling employee teams to understand the reasons for the scores, focusing on questions with low scores and questions where managers and hourly employees have large differences. Out of these discussions, action plans are developed and implemented. We have always found that when the survey is completed and the process is followed, safety performance of the organization improves, usually dramatically. Figure 4 is a control chart prepared by a client, showing the effect of the survey process.

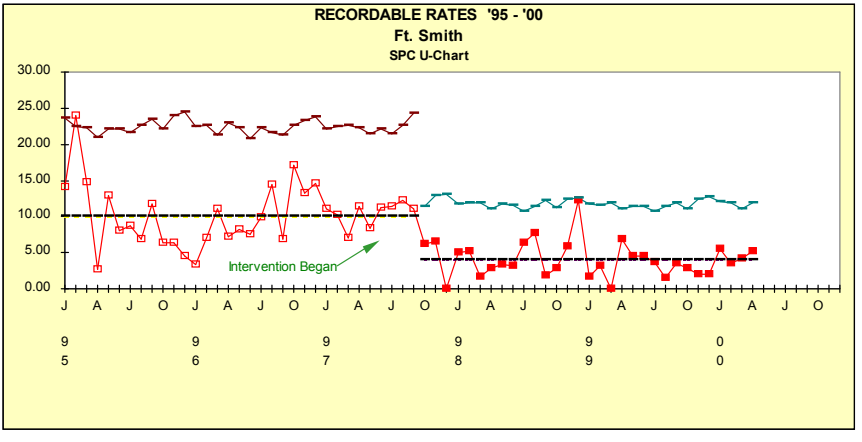


Figure 4

This is a U-chart which is the proper chart for accidents. The dotted line is the process mean, and the upper lines are the upper control limits. There is a process shift after the intervention with a reduction of accident rates of over 50 percent.

With these evaluative principles in mind, let us consider some common measures of business performance. The most universal is

financial statements. This would be a profit and loss statement and balance sheet. Are these reliable? They should be somewhat reliable because the procedures to create them are clearly stated. However, as we shall see later they are a bit less reliable than we might like. They certainly have face validity as they relate to the financial condition of the company. They don't have a great deal of predictive validity, as they don't tell you where the company is going to be next year. They have a little bit of construct validity. For example, they enable you to identify areas of high expense where cuts could be made.

Another common measure is sales volume. Again this has face validity not much predictive validity and perhaps some construct validity as you look at what products are selling, where they are selling, who is buying them, and what kind of margins they are selling at.

Gross margin is frequently a useful measure, although many companies produce financial statements that make it difficult for outsiders to know the true margins. If a company has higher margins than its competitors, it suggests that the company is performing better. This measure has face validity and probably some predictive validity. Although there is some construct validity created by looking at what products and markets are delivering the best margins, this is limited.

Market share is another frequently used measure. Again it has face validity. It probably has some predictive validity. If the company absolutely dominates the market it should be able to maintain that position, at least in the near-term. Historically however we've seen many companies with dominant market share disappear as new technologies, innovation, and better business models, displaced them.

Most companies measure customer satisfaction. It is very important that customers not be dissatisfied, as this predicts decline. However just satisfying customers is not sufficient to maintain or increase market share and profits.

As an investor, I find that none of these measures really satisfies me in terms of predicting the future success of a company. I would prefer some measures that are not available to me. For example, I would like to know about employee morale and employee engagement. Do the employees believe the company is going to be successful? What do the sales people say about the marketplace? Do they expect continuing success or increasing difficulty? What do the technical people say about the company's position? Is the company a leader or is it in danger of falling further behind?

Years ago I was consulting with a company in Silicon Valley. The company was doing relatively well financially but the technical people said that, because the company was falling behind in technology it was in danger of losing its market. In fact that's exactly what happened. The company lost so much money in one quarter that the chairman of the parent company lost his job.

Finally, I would like to know what the company's customers say about the company and what the customers of the company's competitors say about the company. Surveys that only include customers of a company are biased. They are only collecting data from people who like the company well enough to continue to do business with it. It's also useful to find out why customers don't do business with the company, and why they choose a competitor instead.

Were I a large investor, considering buying many millions of dollars worth of stock in the company like Warren buffet does, I think I

would avail myself of some of these measures. They are relatively cheap compared to the amount of money being risked in a substantial purchase.

Much of what is written above is an informal analysis. It is still useful, but far less valuable than a formal analysis. My book, *Measurement Matters*, contains an extensive formal analysis of a number of safety performance measures. If you are using measures that are critical in the guidance of your work, I suggest some formal evaluation.

How using the wrong measure can lead to the wrong action

At the beginning of the primer I mentioned that what gets measured is what gets done. You are at some risk if you are not measuring the right thing. A rather dramatic case of focusing on the wrong measure is the explosion at BP's Texas City refinery on March 23, 2005 which killed 15 workers and injured over 170. When I read about the explosion I sent a letter to Lord Brown who was chairman of BP at that time. I explained a bit about the work I do and how it relates to the type of accident that they had, and suggested they retain my services. In response I received a letter from Lord Brown which came in a rather large envelope. A friend of mine from the UK said that people in Lord Brown's position did not fold their letters. The letter was a polite rejection with the explanation that BP had hired the firm of James Baker, the former Secretary of State, to deal with the situation.

What Baker told them was exactly what I would have told them. In fact they could have simply read my book which would have explained all of this to them. At the refinery, the safety focus was on what you would call personal injuries. These are the things like minor burns, cuts, sprains, etc. that are recorded as accidents in the

statistics that are submitted OSHA. They are used as a performance measure by most companies. Unfortunately the rates of this type of accident do not correlate very well with what we call *process* incidents.

Process safety is related to any production, use, storage, or on site movement of highly hazardous chemicals as defined by OSHA and the EPA. Process incidents can be very large and destructive, such as the refinery explosion we are using as an example. BP had focused on incident rates and had done far too little about process safety. Had they been using our survey system they would have obtained a great deal of information about deficits in their system of process safety management, and would have been under some pressure to deal with them.

While the BP incident is dramatic and very unfortunate, I expect each of you can recall an occasion in which using the wrong measure led to an action that was either unproductive or counterproductive. Since what gets measured is what gets done, you have to be very careful about what gets measured.

How performance measurement can lead to cheating and/or manipulation of the numbers

One of Deming's important insights was the observation that, “whenever there is fear you will get the wrong numbers." An excellent example is found in the book *Freakonomics* by Levitt and Dubner. They studied the Chicago public school system which was using standardized tests to evaluate teachers. They reasoned that some teachers would be induced to cheat, because the consequences of a poor score were quite serious. They further reasoned that the logical way to cheat would be to take some section of answers on the test and mark them all correctly. To mark the whole test correctly

would yield too high a score to be believable. But improving the score by a few questions would be significant. They looked for sequences of correct answers that were statistically improbable and found a number of cases. Using this information they're actually able to get some of the teachers to confess that they had cheated.

In our measurements of safety systems we have found numerous examples of cheating in the recording of accidents.

Figure 5

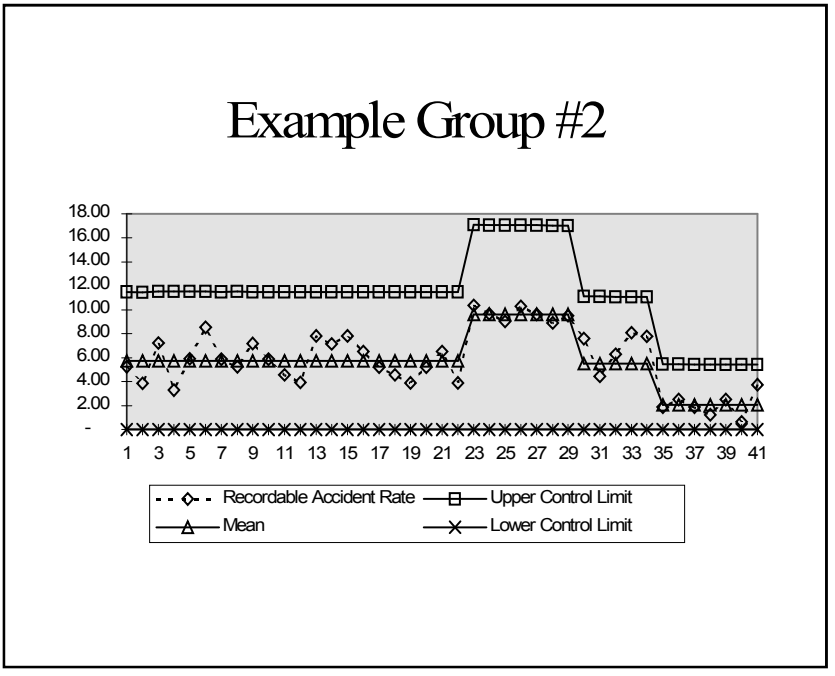


Figure 5 depicts the control chart of accidents in a chemical plant. Note the lack of variability in the second segment of the chart, between period 23 and period 29. The standard deviation of accident rates is proportional to the mean, since accidents are distributed

according to a Poisson distribution. The standard deviation is not computed from the observed variation.

It turned out that the accident rate had gone up and the employees were very concerned about this. What they did to compensate was to stop reporting accidents when the monthly total reached a certain level. Again when confronted with the information from the chart the employees confessed that this was happening.

The next example may represent cheating or simply manipulation of the numbers. Perhaps it is a little of both.

```
DIF Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    60.00 Extremes    (=<-.050)
    12.00       -4 .  000
     6.00       -3 .  0
    18.00       -2 .  0000
    43.00       -1 .  00000000000
    15.00       -0 .  0000
   175.00        0 .  00000000000000000000000000000000000000000000000
   139.00        1 .  0000000000000000000000000000000000
   104.00        2 .  000000000000000000000000
    63.00        3 .  0000000000000000
    39.00        4 .  0000000000
    26.00        5 .  000000
    22.00        6 .  00000
    18.00        7 .  0000
    73.00 Extremes    (>=.080)

 Stem width:       .01
 Each leaf:       4 case(s)
```

Figure 6

What is depicted on this stem and leaf plot in figure 6 is the actual earnings minus predicted earnings for public companies. A minus number means the company fell short. I simply took a sequence of 813 earnings reports from the Bloomberg website to derive this

chart. Ideally a company would hit the prediction. Surpassing the prediction would be better than falling short. What should be expected is a normal distribution, with its center at or near hitting the target.

The actual distribution is certainly centered at the target but it is definitely not normal. Far too many companies are hitting the target. Very few companies missed the target by a slim margin. The probability that the observed distribution is random is vanishingly small, <.0001. What is most likely happening is that creative accountants are finding a way to turn near misses into hits. There many ways that this can be done legally, but if you borrow from the future to look good in the present then it's likely that someday you'll have to pay the piper. Certainly someone looking at this chart should question the reliability of financial reports issued by companies.

The point of all this is that when the consequences of a bad score on a performance measure are potentially very negative, individuals or groups being measured are very likely to cheat or at least manipulate the numbers. In each case, statistical analysis showed that the results did not fit the expected distribution.

The annual merit rating and measuring employee performance

Dr. Deming was strongly against the annual merit rating. He had a number of reasons. Perhaps more than anything he knew that it demoralized too many employees. Moreover he argued that nearly every employee is part of a system and their performance was dependent upon that system. The challenge was to improve the system not challenge the employees. Of course abandoning such a rating left many problems. What employee should be promoted to the next level? Deming proposed that you hire the person you're most comfortable working with. Obviously, in today's environment

that is not going to work. So the problem remains unsolved. While it is necessary to provide for some evaluation of employees it should be understood that this is not easily accomplished and the task of improving the system so that everyone can do a better job is a higher priority.

I have had several dramatic experiences demonstrating how changing the system can dramatically alter the performance of employees.

Larry the sales person. About 30 years ago I was in the senior management of a marketing and promotions company that had a sales force of about 50. Compared to other sales forces in the same industry our team was quite good. Average sales per sales person was in the neighborhood of $500,000 per year and several sales persons were doing in excess of $1 million annually. A young man whom I would call Larry was doing only about $250,000 in sales in spite of the fact that he was bright and energetic. He was considered an underachiever and often treated with some disrespect. He insisted that he could do better if he could sell in a different way. Our sales team had been trained to close an order whenever they were in an office with the customer. He felt he could do much better if, rather than closing the order on first visit, he took some time to develop a plan for the customer and came back for the order on the second call. This would not fly.

Fortunately for Larry a new sales manager was put in charge of the team. He told Larry to go ahead and sell in the way that he wanted to sell. Larry went back to work with new resolve and virtually overnight became the leading salesmen on the team. In fact he became one of the leading salesmen in the industry with sales ranging as high as $3.5 million annually. When the system changed, Larry changed. By the way, his talents were relatively unique, and

through many tried, few of our other salesmen were able to take advantage of Larry's methods.

Mattress Mack takes a gamble. One of the most interesting consulting clients that I ever had was a man known as Mattress Mac. Virtually everyone in Houston knows who he is, because of his TV advertising and his charitable work there. His real name is Jim MacIngvale, and he runs a company in Houston called Gallery Furniture. About 20 years ago, with Dr. Deming's encouragement, I published an article in *Quality Progress* called "Kicking the habit [of poor management]." It described a 12-step method for breaking bad management habits. Mack called me out of the blue said he needed to kick some habits and wanted to come see me.

Mac was devoted to transforming his company along the lines of Deming's philosophy. However he told me that his real problem was that he couldn't get enough high-producing salespersons. For years he had managed to close 42% of the customers who came to his store. In fact the store was extraordinarily successful. Sales per square foot in Gallery furniture were double that of any other furniture store in the United States, and Mac was a wealthy man. But he was determined to raise the closing percentage. And he was determined to transform his company.

Of course he paid his salespersons on commission. He listened to Dr. Deming explain why commission is bad. It encourages things that are bad for the customer like selling the customer more than the customer needs. It discourages cooperation among salespersons.

Mac heard the message. He called me one day and explained that he was ready to end the commission process and put all his salespersons on salary. He asked me to come out and help with this.

Frankly Mac knew exactly what he was going to do and I was not of any help, but I had a wonderful opportunity to observe the process.

No one took a pay cut. Mac put each salesperson on a salary equivalent to their highest commission earnings. He was taking a big risk with an enormously successful company and an excellent sales force. When I saw Mac about six months later he explained what had happened. Pretty soon the highest producing salespersons in the old system quit. However he was now closing over 60% of the people who came to his store. The salespersons that continued to work for Mac cooperated with each other. The customers liked the system much better. And Mac's gamble paid off big time.

Bill Walsh and quarterbacks. My final example of how the system determines performance is about the National Football League. Perhaps I think of this because the NFL season is just beginning as I write this. The story is about the success of quarterbacks under coach Bill Walsh and the system he created. In his system virtually every quarterback that he coached was highly successful. Joe Montana and Steve Young are in the Hall of Fame. Montana arrived as a third-round draft choice. When Steve Young came to Walsh he had 3 wins and 16 losses as a starting quarterback in the NFL at Tampa Bay, with 11 TDs and 22 interceptions. If you are not a football aficionado, I would tell you that these are terrible numbers. Under Walsh, they were both very successful. Montana is frequently deemed the best quarterback in history. Young, whose career was shorter because he was older when he came to Walsh, achieved the third highest passer rating in NFL history. When Montana later went to another team, he was far less distinguished.

Several less famous quarterbacks performed very well under Walsh and less well elsewhere including Guy Benjamin, Matt Cavanaugh,

Steve Bono, and Steve DeBerg. (Apparently, Walsh tended to like guys named Steve.)

The system Walsh designed is called the West Coast offense. Walsh had designed it to maximize the effectiveness of a quarterback who had good mobility, and could throw short passes accurately, but lacked the strong arm for long passes, and to make up for the lack of a consistent running game. Using the system Walsh designed, the 49ers won three super Bowls. After Walsh's retirement, the team quickly won two more super Bowls using Walsh's system. Walsh's coaching tree, men who learned his system from him or his disciples, contains no less than seven Super Bowl winning coaches: George Seifert, who succeeded Walsh in San Francisco, Mike Holmgren, Mike McCarthy, John Gruden, Mike Shanahan, Mike Tomlin, and Tony Dungy. Clearly, Walsh's system made players, and especially quarterbacks, more effective. (He also liked assistants name Mike.)

What these three examples demonstrate is that a change in the system can yield improvements that dwarf the variations in individual performance in the original system. The majority of your time and talent should be focused on improving the system to make everyone a better performer.

Using data

Occasionally in his seminars, Dr. Deming would talk about how managers should deal with data. He exhorted managers to actually use a pencil and paper and "plot the points." He then shouted, "Get the data off the disk." I know well what he meant. I have been involved in the analysis of data for over 50 years. There is no substitute for getting close to the raw data before you perform an analysis. In some circumstances this may not be possible for you. If it is, you should avail yourself of the opportunity. Intuition is a

powerful force. Exposing myself to the raw data sometimes gives me intuitive insight into what is happening. Of course this must then be confirmed by formal statistical analysis.

Even if you can't get the raw data you need to be very careful when looking at highly reduced data. Means and medians may conceal the texture of the data. I will illustrate this with a story from my first consulting assignment. My client was making hard disks. To give you an idea how long ago this was, we were making 20 MB disks. We were moving to 40, which was the cutting edge. The disks were made of aluminum and coated with a cobalt-nickel-chrome magnetic recording surface. The coating was done in a machine called a sputtering machine. The disks would be inserted into the machine and pumps would create a very high vacuum in the machine. Then blocks of the coating substances would be bombarded with high-energy electrons and thereby vaporized into the vacuum. The molecules of the coating substances would deposit on the disks. This created the extremely uniform coating which was necessary for the information storage process. A similar process is used to deposit conductors on microchips.

Of course over time the blocks of coating material, called targets, would be exhausted and would have to be replaced. The problem that was brought to me was that the replacement was taking an average of 36 hours when it should have been taking about 26. The physical work of opening and closing the machine to change the target took about 6 hours. The remainder of the time was required to establish the high vacuum necessary for production. They were losing an average of 10 hours of production on a machine time every 3 to 4 days when targets were changed. Since at the time they could sell every disk they produced, this represented a loss of millions of dollars.

I was given a team of process engineers to solve the problem. They had a solution in mind, and began to describe the modifications necessary to bring the machines into a condition that would enable more rapid target changes. As a businessman, I had some thoughts about the cost of this work, in the hundreds of thousands of dollars for each of the six machines they had. I also had some doubts about whether their solution would really work. Following a hunch I asked them to bring me the raw data of target change times. What I received is depicted in figure 7.

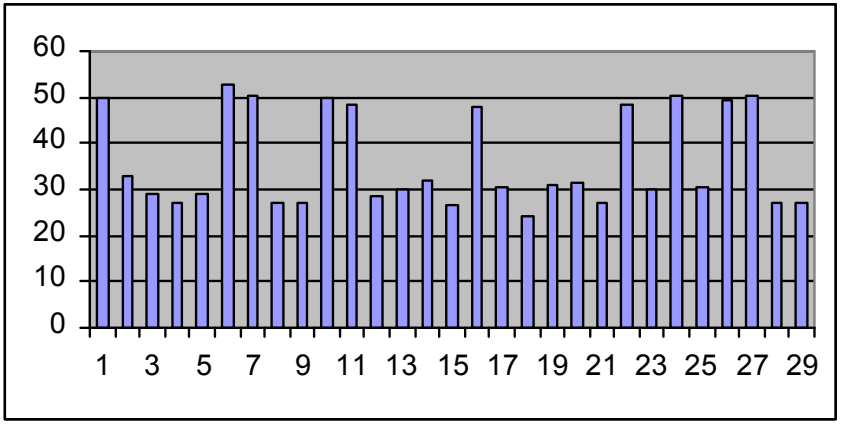


Figure 7

The data are obviously bimodal. There are a number of times in the high 20s and a few times in the range of 50 hours. The obvious next step was to call in the technicians and asked him what happened on those 50-hour occasions. What they told us was that the machine would be opened and the targets replaced. The machine would be closed, and the operation of pumping down to high vacuum would be initiated. At some point in the pumping process they would discover that the machine would not hold sufficient vacuum. It would turn out

that some error had been made in the reassembly, like a washer left out or something. They would have to open the machine, close it properly, and begin the pumping process all over. We asked the technicians what caused the problem during reassembly. They said the problem usually happened when the change operation was passed from one shift to the next. Some piece of information was not passed down properly.

So we offered a simple solution. Target change would be accomplished on one shift only. This meant that the six hours required to open and close the machine would occur on one shift. Sometimes, anticipating a target change that would have to begin late in their shift, the crew would have to initiate the change earlier in the shift to accommodate this rule. The result is depicted in figure 8.

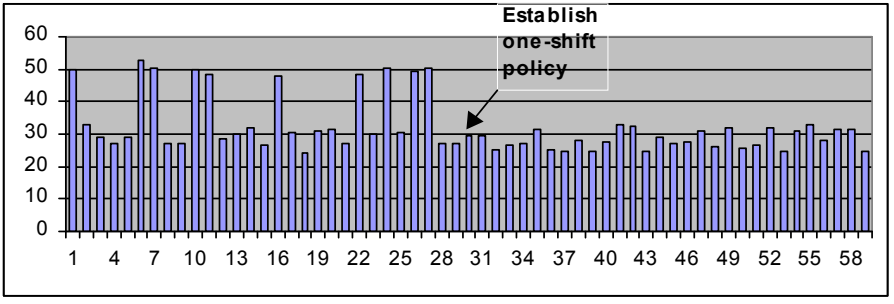


Figure 8

The target change times immediately fell to an average of 28 hours and stayed there. The solution cost almost nothing. An added benefit was that the quality of the product improved significantly. Incidentally, in my experience, shift changes are a source of many problems, including serious safety problems.

For example, in the Piper Alpha disaster in 1988, an explosion and fire occurred on a North Sea natural gas production platform operated by Occidental Petroleum. One of the important causes was a failure to effectively pass a critical piece of information from one shift to the next. The ultimate result was 168 lives lost, one of the largest industrial accidents in history.

Measuring complex systems

When I was in graduate school in the late 1960's, one of my Professors, Dr. Phil Teitlebaum, explained the importance of using more than one measure when you are assessing complex systems. This stuck with me over the years, and grew in significance through the early years of my professional career. When attempting to measure the performance of a business or an individual, on virtually any dimension, you are measuring a complex system. There is no such thing as a perfect measure. As we noted before, Deming stated that "There is no true value of anything. There is a measurement method and a result." Consequently, it makes sense to use more than one measure. Each measure that you use should have some reliability and validity. I cannot tell you exactly how to combine the measures. It depends entirely on the circumstance.

One thing I look for is where the measures diverge. Attempting to understand why they diverge is likely to provide important insight. For example, years ago the marketing and promotion company I ran grew to about 110 employees. Our customers loved us, so customer satisfaction was very high. I have only seen two companies with higher scores. Our employees were highly engaged. We had virtually no turnover and very high employee morale. However, our financials were not very good.

Profits ranged from 1% to 3% of sales. Because of the low profit, the company had a weak balance sheet. Was the company performing well? In my opinion, no, it was not performing well. My conclusion from the data was/is that the company had a bad business model. In the economic slowdown that followed 9/11, four of the five largest California-based companies in our industry that were using our model, including ours, failed. At the time we were attempting to change the business model, but we were too late.

Another example of a discrepancy would be Gallery Furniture which originally had very strong financial performance and engaged employees but likely had low customer satisfaction. I never saw formal data on the customer satisfaction, but expect this is so from my conversations with several friends who had visited the store. A rational hypothesis to explain this discrepancy is that the compensation model was faulty. Certainly when the new model was introduced, customer satisfaction improved and so did profits.

Some advice for going forward

If I were to pick one piece of advice related to performance measurement, I would say that you should not take numbers that are given to you on important issues at face value. How were the data collected? Is there any evidence of reliability and/or validity? Are there any other data on the same issue? Do the other data support the present finding, or are there discrepancies? How strong is the possibility that the data are being manipulated or fudged? What motivation might there be to do this? Are the data being fully utilized, or just superficially analyzed?

If you are using data that are important to your operation, treat it like you own it, not like you are renting it.

References

Carder, B and P. Ragan, *Measurement Matters: How Effective Assessment Drives business and Safety Performance*, Milwaukee: ASQ Quality Press, 2004.

Deming, W. E. *Out of the Crisis*. Cambridge, MA: MIT Center for Advanced Engineering Study, 1986.

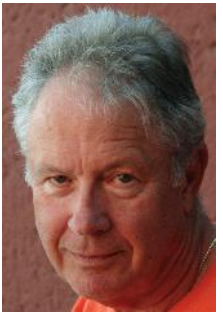
Levitt, S. D. and S. J. Dubner, *Freakonomics: A Rogue Economist explores the Hidden Side of Everything,* New York: William Morrow, 2005.

Lewis, M. *Moneyball*. New York: W. W. Norton and Company, 2003.

Acknowledgement

First of all, I must thank Patrick Ragan who was my partner in all of the safety work described here. Pat is currently an executive with Bayer CropScience. His sharp mind and extensive knowledge of safety have driven much of this work over the nearly 20 years that we have worked together. I would also like to thank my colleagues in the leadership team of ASQ's Human Development and Leadership Division for their support, and particularly Bill Barton for his thoughtful reviews of this manuscript.

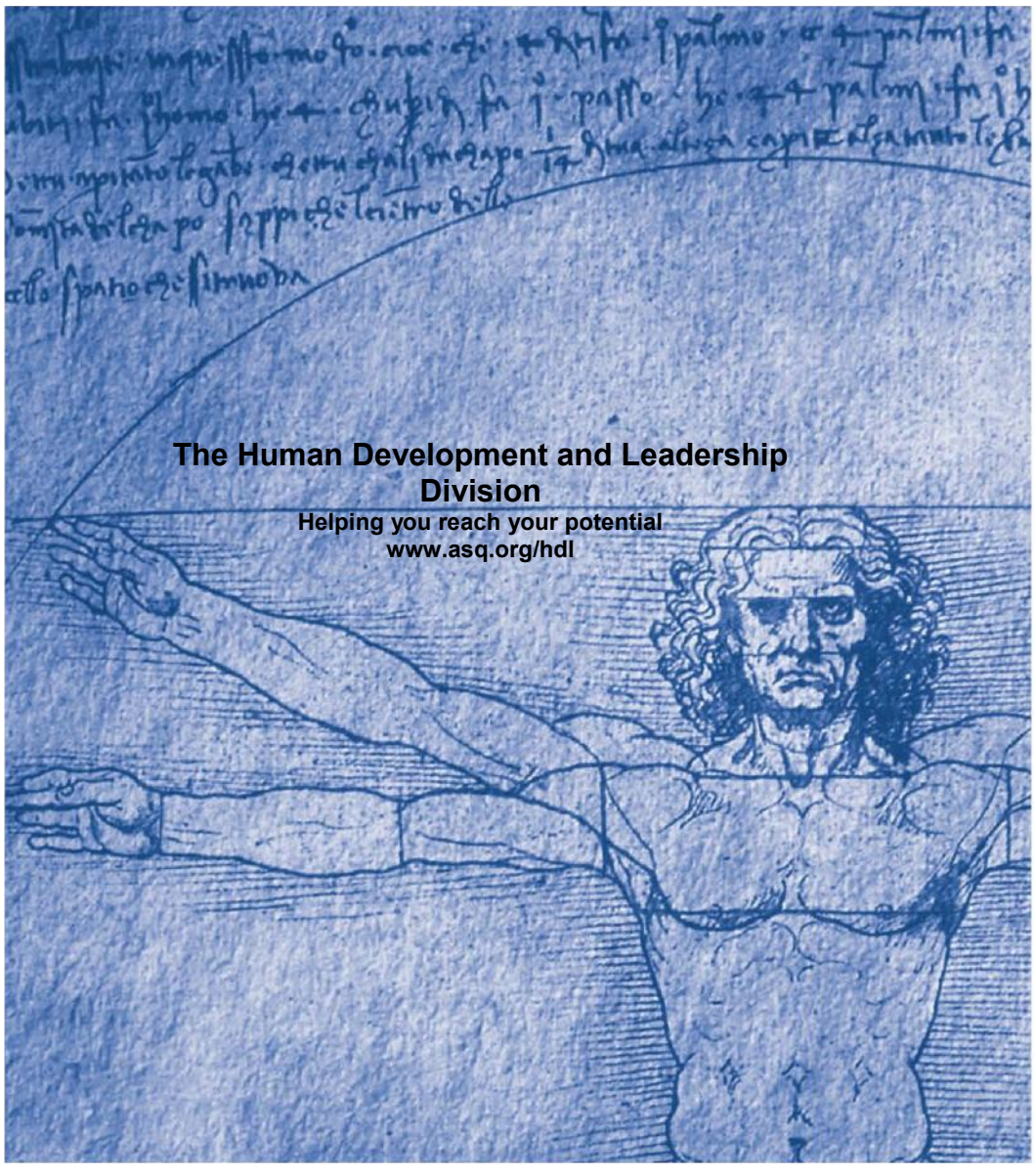
About the Author, Brooks Carder, PhD



Brooks is a leader in the application of the disciplines of Quality to the improvement of safety performance. His work focuses on the use of scientific assessment to enable the design and implementation of successful improvement programs. He and his colleague, Patrick Ragan, authored the chapter on "Benchmarking and Performance Measurement" in *The Safety Professional's Handbook*, and a book, *Measurement Matters; How Effective Assessment Drives Business and Safety Performance*, published by ASQ Quality Press.

Brooks' interest in measurement was inspired by his contact with Dr. W. Edwards Deming in the 1980's. The webinar material is the result of Brooks' subsequent 25 years of experience with measurement, beginning on the factory floor in the Silicon Valley, and ranging from the measurement of customer satisfaction for Fortune 50 companies, to the measurement and improvement of safety system performance in large chemical plants.

Brooks lives with his wife Fran, in Del Mar California. He is an avid bicyclist, golfer, and cook.



The Human Development and Leadership Division

Helping you reach your potential

www.asq.org/hdl

ASQ

600 N. Plankinton Ave.
Milwaukee, WI 53201-3005 USA
USA and Canada: 800-248-1946
Mexico: 001-800-514-1564
International: +1-414-272-8575
Fax: +1-414-272-1734
www.asq.org

x0008f92n9

X0008F92N9

Measuring Performance...nd Damned If You Don't

New