

SECTION 1 RISK ASSESSMENT AND HAZARD CONTROL

LEARNING OBJECTIVES

- Understand the concepts of reliability and validity and how they are critical to the evaluation of any measurement process.
- Be able to apply these concepts to performance measurement.
- Conduct better performance measurements based on an understanding of research that evaluates the reliability and validity of incident-based measures, audits, and surveys.
- Understand the limitations of performance measurement and the hazards of incentives for performance that can lead to manipulating the results.
- Be able to identify instances of manipulation of results.
- Develop the ability to design, conduct, and evaluate a productive benchmarking study.

BENCHMARKING AND PERFORMANCE CRITERIA

Brooks Carder and Pat Ragan

PERFORMANCE MEASUREMENT is a fundamental step in risk assessment. In a stable system, performance will remain the same until the underlying process changes, so a measure of current performance constitutes an assessment of future risk (Deming 1982, 1993). Performance measurement and benchmarking are both methods that can assist in hazard control, by revealing opportunities for process improvement.

THE RELATIONSHIP BETWEEN PERFORMANCE MEASUREMENT AND BENCHMARKING

Performance measurement and benchmarking are obviously intertwined. Merriam Webster's online dictionary (www.m-w.com) defines *benchmarking* as "the study of a competitor's product or business practices in order to improve the performance of one's own company." However, the term derives from the noun *benchmark*. The definition of a benchmark includes "a point of reference from which measurements may be made" and "something that serves as a standard by which others may be measured or judged." Performance measurement is usually not very meaningful unless there is a benchmark for comparison. If you are asked how fast someone is going, and you get an answer of 100 miles per hour, you would think that was extremely fast on a bicycle, fast in a car, not very fast in a racing car, and extremely slow in a jet plane. To the extent that performance measurement is evaluative, there must be an explicit or implied benchmark.

On the other hand, to the extent that benchmarking represents an attempt to improve performance, it is necessary to find benchmarking partners that have excellent processes and excellent performance (Camp 1995). The objective is to identify and

implement the processes that lead to superior performance in other companies. Thus, benchmarking cannot be done effectively in the absence of good performance measurement. Keeping these interrelationships in mind, the chapter first addresses performance measurement and then benchmarking.

PERFORMANCE APPRAISAL

Defining Performance Appraisal

On the face of it, performance appraisal in safety should be very simple. One can simply count injuries, deaths, and property loss. In reality, however, the problem is very difficult. The following problems arise:

- Some industries and activities are inherently more hazardous than others.
- Over a short period of time or with a relatively small population, the inherent variability of these counts is high, making judgment based on the numbers very inaccurate.
- In an environment where a major disaster could occur, such as with an airline, a chemical plant, or a refinery, assessing the likelihood of a major event should be a top priority. These are so rare that, thankfully, in most sites, there is nothing to count, even though the danger may be high.

Dictionary.com defines *safety* as freedom from danger, risk, or injury. Conditions are easily conceived in which there is no history of injury but great risk of future injury. Of course this appears to be the case with shuttle flights up to the time of the Challenger and Columbia disasters. Although there was no history of injury, the engineers working on the flights estimated the probability of the loss of a vehicle in the range of 1 in 100 (Feynman 1999).

Ideally, a measure of performance would tell us the level of freedom from danger, risk, and injury. The measure would not be a picture in the rearview mirror, but rather an accurate forecast of future expectations, so long as the system is not changed. Many readers may believe that incident counts are indeed an accurate forecast of overall safety. But the

Section 1: Risk Assessment and Hazard Control

available evidence indicates that this is not the case. Part of the problem lies in the lack of reliability of incident counts because the standards for OSHA-recordable events can vary from company to company, and even from day to day in the same company (Carder and Ragan 2004). An article in *Professional Safety* describes how recordable counts can be altered through "medical management of injuries" (Rosier 1997). This practice is widespread and is a significant limitation to the reliability of accident counts (Carder and Ragan 2004). Another problem is that accident counts have proven to be a poor predictor of catastrophic events (Manuele 2003, Petersen 2000, Wolf and Berniker 1999).

Objectives of Performance Appraisal

An important objective of performance appraisal is to provide information to guide improvement efforts. Another is to track the effectiveness of improvements that are implemented. This is the plan-do-study-act cycle described by Deming (1982). Closely related to this is the need to evaluate the performance of managers and to provide guidance for establishing reward systems.

Hazards of Performance Appraisal

The first question to be asked is whether an accurate meaningful assessment can really be made. This chapter suggests that one can indeed make a useful assessment of the safety performance of an organization or subunit. The second question is, whose performance is being appraised? A safety manager in a plant is part of a system. He or she usually has very limited control over the larger system. The system includes such practices as hiring policies, education and training, manpower decisions, budgets, capital expenditures, and much more. All of the things mentioned have an impact on safety. Although one can measure the performance of the system, it is much more difficult to measure the performance of individuals working in that system. Deming (1982) argues continuously and eloquently that attempting to evaluate the performance of individuals working in a complex system is a waste of time. Nevertheless, it is unlikely that business

Benchmarking and Performance Criteria

will move away from this anytime soon. However, the reader should be aware of the limitations of such evaluations when they are used.

Consider the following actual case study: Many years ago a marketing company had a young man in sales who was very bright and energetic. However, his performance in sales was continuously disappointing to his superiors. He was labeled an underachiever and, in private conversations, much worse. He constantly asked his managers to be allowed to sell in a different way and was constantly told that the company had a system of proven success and that he should sell exactly as he was told. The sales rep wanted to uncover marketing problems that confronted the customer and return to his office to prepare a solution. The solution would be presented to the customer on a subsequent visit. He was told that he needed to present a solution and close the order on the initial visit, like all of his successful colleagues. One day the management system changed, and his new manager told him to go out and sell in the way he wanted. Within a year he was the company's top salesperson and a leader in the industry. Up to this time, the company had considered a \$10,000 order to be very large. After the system change, this rep wrote orders as large as \$500,000, at higher margins. Changing the system dramatically changed his performance. At best one can measure only the interaction between an individual and the system in which he or she works (Deming 1982).

One of the worst risks of conducting a performance appraisal is that when rewards are based on that appraisal, it can provide an incentive to game the system. Levitt and Dubner's recent book (2005), *Freakonomics*, describes, in considerable detail, a number of cases of how reward systems lead to cheating. This is not an accusation that managers commit fraud in order to secure bonuses. Although this has happened, as evidenced by the accounting fraud convictions in the cases of Enron and World Com, it is hopefully rare. However, there is an inherent conflict of interest in basing the pay of a person who is measuring something on the result of that measurement process. An example of this kind of manipulation is seen in Figure 1 (Carder and Ragan 2004).

Figure 1 shows a control chart of recordable accidents for Group 2, one of several manufacturing units in a large plant. Each point on the x-axis represents one month. The y-axis is the rate of recordable accidents. There is an upward shift around months 23 to 28. This shift illustrates a process shift in the wrong direction with seven consecutive points above the previous mean. For rules of interpreting control charts, the reader can refer to Nelson (1984). Although the output of a stable process will vary, certain patterns in the variation indicate that the process has changed, indicating a special cause. Special causes need to be investigated. Some special causes indicate that there is something wrong with the measurement process. On closer examination, the next series of points is very close to the new mean. According to the rules of control charts (Nelson 1984), a special cause requires the finding of fifteen points within one standard deviation of the mean. In this case, this condition is not met because there is another process shift at month 30. However, the points between months 23 and 29 are much closer to the mean than one standard deviation, suggesting the presence of a special cause. Upon investigation, it was found that because of the upward shift in the incident rate, managers in Group 2 put extreme pressure on the group to hold down the accident rates. In their zeal to turn the trend, they did not realize they had gotten exactly what they asked for. People stopped reporting accidents. Accidents happened at the rate expected for the process; they simply

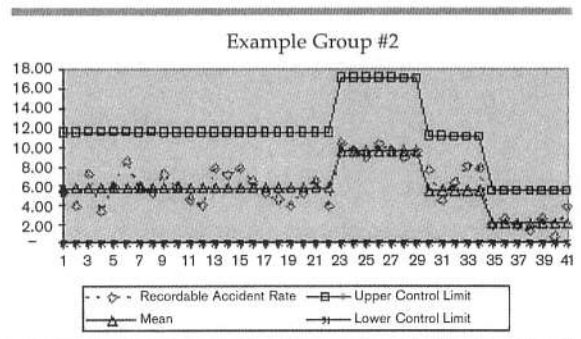


FIGURE 1. Control chart of accidents by month, showing process shifts

did not report those they could hide or classify as not being recordable. As the number of accidents in the month increased, pressure to not report also increased. Part of the reason the rates did not drop farther was that some accidents were just too serious to hide. This actual case is a clear illustration of manipulating the data to achieve an outcome and shows the value of using effective measurement tools to identify when this happens. Here the safety managers knew that the existence of nonrandom patterns indicated that something was wrong with the measurement process. Such patterns are frequently an indication that someone is manipulating the data (Deming 1982). The safety managers used that understanding to focus attention on the causes of the statistical anomaly and acted to correct it.

In order to guard against the risks involved in assessing performance, one must have an understanding of the principles of measurement. Application of these principles is critical to a meaningful assessment of performance. Incidentally, the application of these principles to Enron, World Com, and Health South would probably have revealed the problems early in the game. For example, Banstetter (2002) points out that several analysts who looked carefully at a variety of measures of Enron's performance were able to foresee serious problems. While revenues were growing, net margins were shrinking. Potential risks were obscured by keeping them off the balance sheet. Also, senior management was dumping a lot of stock.

Principles of Measurement

The quality of a measure is determined by its reliability and validity. No measure is perfect, and all measures have limitations on their reliability and validity (Deming 1982). Deming frequently illustrates this point by referring to measurements of the speed of light, which is an important constant in physics. He charted the variation of this measurement over time. Because the speed of light is assumed to be constant, the chart shows the variation in the measurement. Another illustration he used is variation in the census, depending on the method used. He noted that there is "no true value of anything. There is a measurement method and a

result." If the method changes, it is likely the result will change.

Reliability of a Measurement

Essentially, *reliability* refers to the repeatability of a measure. For example, there is a particular golf course in central California at which the yardage markers appear to be unreliable. The markers that signify 150 yards to the green appear to be out of place on the eighth and ninth holes, reflecting a distance of closer to 160 yards. This may be explained by the assertion of one of the course employees that the persons who put in the markers used a 150-yard rope to make the measurements. The rope was nylon and got wet as they proceeded. The wet rope stretched, yielding longer distances on the later holes.

One could use many methods to measure the course, ranging from pacing off the distance to using a steel cable, laser, or global positioning system. Measuring a distance repeatedly with each method, would likely yield a spread of numbers around an average in each case. This is called *spread variation*. Variation is quantified by computing the standard deviation. The standard deviation is a statistical estimate that quantifies the variability of a measure. The greater the standard deviation, the greater the variability of the measure. The variability of measurements from the laser, presuming it was working properly, would be much smaller than the others. The wet nylon rope would very likely have the most variation. One would conclude that the laser was more reliable. There is no such thing as a perfectly reliable measure. All measures will show variation.

In some cases reliability is assessed by looking at the variation between observers. For example, audits are scored based on the judgment of an auditor or audit team. To judge the reliability, ask whether a different audit team, unaware of the initial team's evaluation, would give the same or a similar score. In practice, have two auditors conduct audits of a number of sites. Each auditor would be unaware of the scores given by the other. Then compute a correlation coefficient between the scores given by the two auditors. A correlation coefficient assesses the degree to which one measure predicts another. In this case one is assessing

whether the score assigned by one team predicts the score of the other. A high value, meaning that the scores of one team are good predictors of the scores of the other, would indicate good reliability. A very low or even negative correlation would mean that the process has no reliability at all, in this particular test. It is still possible that one of the auditors is very accurate. However, there are two problems with this: (1) one can't know which auditor was correct, and (2) even if that could be determined, the process would be dependent on one person's judgment. In this case the measure is neither reliable nor useful. In order to develop a reliable and useful measure, one might attempt to clarify and better define the criteria and methods, retrain the auditors, and test again for reliability on a different set of plants.

No matter how reliable a measure might be, it will still have variation. If two auditors consistently report exactly the same score for the audit of a plant, management should question whether they are really operating independently. As mentioned previously, anomalies in variation often signal a problem with the measurement process. If a measure has no variation at all, one is not looking closely enough, the gauge is broken, or the numbers are being manipulated by the observer.

Reliability is not a property of the instruments, but of the entire measurement process. This includes the tools, the instruments, the procedures, and the people. Subjective judgment can be very reliable in some cases, and measurement with the finest instruments can be unreliable if the process that uses these instruments is flawed.

To repeat, any measure will vary. If the measure does not vary, or if the pattern of variation is not normal, an investigation is warranted. Assuming there are no anomalies in the variation, then the less the variation, the higher the reliability. For any particular purpose, there is a level of reliability that is acceptable for the task. When carpeting a room, one can use a tape measure, but not one that is elastic. When measuring length in order to construct a complex optical system, the laser might be required.

In the safety area of a business, it must be realized that all measures have reliability limitations and that all measures are subject to manipulation. It is impor-

tant to understand the limitations of each measure and to take these limitations into account when making decisions based on the measurements.

It is important to realize also that a measure with high reliability may still be worthless if it does not convey anything useful about the what one is trying to measure. Just because a measure is reliable does not mean it will help management take effective action. This leads to the concept of validity.

Validity of a Measurement

Validity relates to whether one is measuring what he or she wants to measure. When measuring the width of a room, the question of validity usually does not arise. When measuring a complex process such as aptitude to perform well in college, or the ability of the safety management system in a plant to prevent future loss, validity becomes a serious question. Scientists (Cronbach and Meehl 1955) generally define three categories of validity: content-related validity, criterion-related validity, and construct-related validity.

CONTENT-RELATED VALIDITY

Often called *face validity*, content-related validity asks whether the content of the measurement process is, on its face, related to the purpose of the measurement. A good example is found in safety audits. If a question asks whether employees use personal protective equipment on the factory floor, that question has face validity. If one asks whether employees go out for beer together after work, that lacks face validity because nothing in the content appears to have anything to do with safety. However, the question might have criterion-related validity and construct-related validity. It could turn out that when employees have close personal relationships the plant is safer, and that going out for a beer (or bowling, or to church, etc.) after work is evidence of such relationships. Of course, this is not an assertion that this is actually the case.

CRITERION-RELATED VALIDITY

This is sometimes called *predictive validity*. It deals with whether one measure correlates with other measures that could be called criteria. For example, the SAT test

is an attempt to measure the likelihood that a student will succeed in college. Obviously the criterion here is college performance. Yale University has used the SAT for many years. Although they admit it is not a very good predictor, they also say that it is the best they have. This is because different high schools have very different criteria for grading, so that an A in one might be equivalent to a C in another. Over the years, the Pearson correlation between SAT scores and Yale grades has run in the range of 0.2–.03. This is statistically significant, meaning that the SAT does indeed have criterion-related validity. However, this level of correlation means that, at best, what is measured by the SAT is accounting for no more than 9 percent of the variation in college grades. (Squaring the correlation coefficient of 0.3 gives us the percent of variation accounted for, 0.09.) The other 91 percent is presumably accounted for by other things, such as motivation, the quality of the student's secondary education, the difficulty of the courses chosen at Yale, luck, or any number of other variables each student must face and overcome to "make the grade" at the university.

When dealing with large populations, it may make good economic sense to use measures such as the SAT, which have relatively low criterion-related validity. However, individuals who are negatively affected by such measures will always have a pretty good argument that the measurement was unfair to them.

It is also important to realize that the criterion is arbitrary. After all, the success of a Yale career should not be measured by grades. Yale and most other universities are interested in producing good, productive citizens and leaders. Does the SAT predict that?

In the safety field, injuries and monetary losses are certainly useful criteria against which to test other measurements. However, they are not the only possible criteria, and they may not be the best criteria. Ultimately, one would like to know the ability of the safety management system to prevent future accidents and losses. Although a burned finger may be of some concern in a chemical plant, it is trivial in comparison with the release of a toxic chemical that could injure or kill thousands of people. Because catastrophes are fortunately infrequent, they are inconvenient to use as criteria in a validation study. An excellent

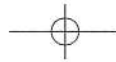
safety measure would enable the prevention of catastrophic events. Minor incidents such as burned fingers are associated with cost and suffering, but they are neither equivalent nor directly related to catastrophic events (Manuele 2003, Peterson 2000).

Because criteria are usually somewhat arbitrary, and because there is often no single, ultimate criterion, it is best to use several criteria when attempting to establish criterion-based validity. This not only will make it more likely that validity will be established, but also will likely increase understanding of the measure being tested and the results attained. This leads to the concept of construct-related validity.

CONSTRUCT-RELATED VALIDITY

Construct-related validity goes to the understanding of what is being measured. In the measurement of safety performance, there has been little work on construct validity. Carder and Ragan (2004) used a reliable and valid safety survey to measure performance. They found that the critical constructs measured by the survey were (1) management's demonstration of commitment to safety, (2) education and knowledge of the workforce, (3) quality of the safety supervisory process, and (4) employee involvement and commitment. Coyle, Sleeman, and Adams (1995), working with a different survey, identified a similar set of constructs.

Safety professionals have been inclined to take the measures they are using for granted as a result of their content-based validity. If their intuitive feeling is that the measure is valid, they use it. They look to the content of their measurements to determine exactly what is being measured. There is a serious limitation in this practice. The most obvious case is the way incidents are counted. A minor cut or burn is recorded and investigated. On the other hand, many more important events, such as a chemical reaction going temporarily out of control, are often neither recorded nor investigated. The assumption, based on faith rather than evidence, is that the burn and the control of the chemical process are the same thing. Although they may be related, given they are both outputs of the overall management system, the evidence indicates that they are not the same thing (Wolf and Berniker 1999, Manuele 2003, Petersen 2000).



A better understanding what is being measured, leads to better judgment of what management actions are suggested by the measurement (Carder and Ragan 2004).

Managers prefer measures they believe to be highly reliable, in spite of the lack of any evidence regarding the validity of those measures. Many managers believe that surveys and interviews are of doubtful reliability and that the measurement of incident rates is quite reliable. In fact the data show that surveys and interviews can be very reliable, whereas the recording of incidents is frequently unreliable. In a recent book (Carder and Ragan 2004), we describe numerous examples from our own personal experience of how incident rates can be unreliable. More important is the limited ability of incident rates as a measure to enable the prevention of future catastrophic loss (Wolf and Berniker 1999, Manuele 2003, Petersen 2000). This is an important limitation of the construct-based validity of incident-rate measures.

A measure with moderate reliability and high construct-related validity is much preferred over a measure with high reliability and little or no evidence of construct-related validity. The latter measure may be very accurate, but it does not provide anything useful to guide management actions. If it is used to guide action, the effort will likely be wasted. Because it has little construct validity, it tells very little about the process one is attempting to improve. It would be like using a map of Ohio to drive in Pennsylvania.

Usability of a Measurement

Usability refers to the ease and cost of the following:

- collecting the data
- analyzing the data
- communicating measurement results
- using results to devise action plans

COLLECTING DATA

In 1994, a safety survey (Carder and Ragan 2003, 2004), which is discussed in detail later on, was developed. The survey was conducted in more than 50 plants of a major chemical manufacturer. In the previous year, the company had established a manufac-

turing strategy team (MST) to evaluate the quality of the management system in the same plants. When the plants' survey scores were compared with the MST scores, the correlation coefficient was -0.58 . The correlation is negative because on the MST, lower scores indicated better performance, whereas with the safety survey, higher scores indicated better performance. This correlation indicates that the two processes were measuring many of the same things. However, the MST process required two to three highly trained and experienced staff members to conduct several days of interviews at each plant. The survey process required the employees to fill out a simple yes/no survey that took 20–30 minutes, usually during an already scheduled safety meeting. The scoring of the surveys was automated, and the analysis of the resulting data was relatively simple. Moreover, the survey had high reliability. The reliability of the MST process had not been evaluated. Thus, as a measure of performance, the survey was less expensive and simpler to implement than the MST process. It had much higher usability. Actually, the strong correlation between the survey and MST represented a validation of the MST process because the survey had been validated against other criteria as well.

An important lesson taken from this experience was that a good management system is likely to yield good safety performance. A tool that is used to measure the quality of the safety-management (accident-prevention) system also gives a good evaluation of the quality of the overall management system.¹ The survey was a simple, quick, and inexpensive tool that provided insight into how well a group was being managed. The alternative approaches to measuring the overall quality of the management system had a cost that was orders of magnitude higher, required much more effort, and produced less statistically supportable results. The survey tool is in the public domain, as are descriptions of how it can be used for measurement and for process improvement (Carder and Ragan 2003, 2004).

ANALYZING DATA

Managers often want to reduce performance evaluations to numerical data. This simplifies comparison

