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. The benchmark is needed even if it is only the benchmark of change for the measure from its previous position(s).

On the other hand, to the extent that benchmarking represents an attempt to improve performance, it is necessary to find

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 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 appraisal 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, at 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). In fact, hindsight suggests that this is a very good estimate, with 2 vehicles lost in 130 flights as of February 8, 2010.

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 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). While legitimate and widespread, this practice introduces a great deal of variability and places a significant limitation on the reliability of accident counts (Carder and Ragan 2004). Adding to the problem is that accident counts have proven to be poor predictors of catastrophic events (Manuele 2003, Petersen 2000, Wolf and Berniker 1999).

Rosenthal and his colleagues (Elliott et al, 2008) have addressed the problem of predicting major accidents for a number of years, and have done the most exhaustive study of the relationship between incident rates and the occurrence of what they call “low probability-high consequence” (LP-HC) events. For incident rates they used OSHA recordables, and for LC-HC events they used events reportable to the EPA under the Risk Management Program of the Clean Air Act. In fact, the facilities with the lowest rates of LP-HC events tended to have the highest recordable rates. However, this was because facilities such as chemical plants and refineries tend to have low incident rates but high risk, while facilities like poultry processors have high incident rates, but low risk of LP-HC events. When they controlled for this risk, they found a trend toward a positive relationship between incident rates and HP-HC events, but it was “far from statistical significance.”

While the reader will not be left without some notions of a solution to the problem of performance measurement, the problem of assessing the ability of the safety management system to prevent major events remains unsolved. However, it does appear clear that achieving a low incident rate should not lead to confidence that the system is effective at preventing LP-HC events. Elliott et. al. point out that in the Texas City
refinery explosion, which killed 15 workers, BP had achieved a low incident rate. The Baker Panel Report (Baker et al 2007) went so far as to suggest that reliance on injury rates might even mask or distract an organization from measures and actions that would identify and avoid these LP-HC incidents. They note that BP had emphasized personal safety but not process safety and “mistakenly interpreted improving personal injury rates as an indication of acceptable process safety performance at its US refineries.”

**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 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
those who are impacted by it or who need to respond to it is critical. In the example being followed, processes such as the MST evaluation sometimes have a usability advantage. If a manager has a high degree of confidence in a staff member, then reports by the staff member are likely to generate action. Of course there is a problem with this, in that the usability of this kind of report is dependent on the relationship between the manager and the staff member. It may not always be possible to find the right staff member for a particular evaluation. With survey data, experience indicates that when the reliability and validity of the survey are carefully explained to the manager, action is usually generated (Carder and Ragan 2004). This should be done carefully and thoroughly by someone who has a good understanding of the survey process.

USING MEASUREMENT RESULTS

Measurement results should be used to devise effective action plans. This is the last consideration, but a critical one. For example, counting accidents tells one little about what to do. Often managers see an accident rate that is too high and tell their subordinates to try harder to be safe. This is not useful. Deming (1993) describes exactly this kind of event, and we have seen it ourselves many times. Deming recommends “looking into the process” that produced the accidents, rather than exhorting people to do better. A way of looking into the process is to find the true root causes of accidents using an effective investigation process. Then one can work to eliminate those causes, thereby reducing the potential for many future accidents.

If there are two methods with relatively equal reliability and validity, it makes sense to employ the one with the highest usability. However, although usability is important, remember that if a measure has very poor reliability or validity, the strong usability is worthless. It is like a man who is found one evening looking for his car keys on the street. A passerby questions him, finally asking if this is where the keys were lost. The man tells him he actually lost them in the middle of the block. The passerby asks why the man is looking for them on the corner. The searcher replies that he is searching at the corner because the light is better.

EFFECTIVE USE OF AVAILABLE MEASURES FOR PERFORMANCE APPRAISAL

The sections that follow examine some widely used types of measurement for reliability, validity, and usability, and ultimately for their value as measures of safety performance.

Incident-Based Measures

An almost universal measure of safety performance is based on incidents: the recording of accidents and the investigation of their causes. Since 1970 OSHA has required companies with eleven or more employees to maintain a record of accidents and injuries. Many other countries use incident-based measures with different incident definitions and different normalizing factors. In the United States the number is based on the number of incidents per 200,000 exposure hours. In Europe and many other countries the number of exposure hours used is 1,000,000.

The mere fact that accident counts are the fundamental method by which the government measures safety suggests that they can be measured very reliably. Although there are frequent and very articulate complaints that incident-based measures are not very helpful in process improvement (Petersen 1998), there is rarely a question about their reliability or validity. But, in fact, an examination of reliability and validity suggests that there are serious limitations to performance measures based on the counting of incidents. Many of these limitations are discussed in the following sections. For a more extensive treatment of this issue, see Carder and Ragan (2004).

Reliability of Incident Rates

There are two sources of limitation on the reliability of incident rates: variation in interpretation of the criteria for recording an incident and inherent variation in the statistic itself.
VARIATION IN RECORDING CRITERIA

Recording incidents is relatively complex. Studies of the reliability of classifying events as recordable or not recordable are lacking. However, there are many examples of stretching of the criteria:

1. Use of over-the-counter ibuprofen instead of prescription dosages. If the prescription is not used, then recording the incident is not required. Many doctors will cooperate with this approach. This cooperation is rationalized by the assumption that this is what the employee or employer wants, that it provides the same relief for the injured employee, and that it is less expensive than prescription medication.

2. First aid given for increasingly serious injuries that could have or, in some cases, should have warranted medical treatment. An article in Professional Safety (Rosier 1997) actually recommends setting up first-aid stations to “prevent the accident from falling into the OSHA recordable category.”

3. Liberal interpretations of preexisting conditions are made to avoid recording an incident or to count a case as one case instead of two. Sometimes the first injury or illness case will have been in a different year, so even if the second is counted, the first incident is not included in the measures for the time being considered. It can be recorded with no negative effect because most companies do not factor in such historical changes.

4. Classification of more events as not being work-related. The pendulum swings from taking the employee’s word for a case being work-related to requiring the employee to prove the case is work-related beyond any doubt.

5. Employees being offered full pay to work at home when their injuries prohibit their working at their normal workplace or traveling to and from their workplace. The requirement is that they go along with the story that the injury did not result in lost work.

6. Employee job definitions being used to define work relationship. In one case, an employee fell from a scaffold, breaking both wrists and suffering multiple other injuries. The injuries resulted in a hospital stay, but the case was initially not included on the injury log because it was argued that the employee was not doing his regular job. It seems this was an infrequent task that was not in his regular job description.

7. Manipulation of medical diagnosis. For example, an employee was cut and received fifteen stitches. After consulting with a physician, the employer argued that the stitches were “cosmetic,” not “therapeutic,” and therefore the case was not counted.

These are only a few of the many approaches used to avoid counting. It makes one feel there may be some truth in an old joke about an accident: An employee fell from a rooftop, and his scream drew the attention of his supervisor. It was lucky for the company that he screamed because it saved their twelve-year no-lost-day case record. It seems the supervisor was able to fire him for a safety violation before he hit the ground and was injured.

Although these might be extreme examples, they illustrate the difficulty in comparing the incident rates of two companies and being certain that the one with the lower rate truly has better safety performance. Petersen (1998) notes that accident rates do not discriminate between good and poor performers. Consider the plausible scenario where one company has a low accident rate because it underreports, and another company has a relatively high accident rate because it sees the reporting of accidents as an opportunity for investigation and improvement. The company with the higher reported record may actually have the better safety management system.

The problem noted here is not so much that employers are scofflaws and do everything they can to avoid the record-keeping rules out of contempt or malice. Frequently managers do not even know when “marginal” reporting is taking place. In almost every case of which we are aware, the companies involved