CHAPTER ONE

Measurement, Evaluation, and Research

Feedback for Decision Making

Carl Binder

In his elegant little book devoted, not to measurement and evaluation, but to the essence of sustainable performance improvement, Esque (2001, p. 18) states that:

"In its simplest form, managing work consists of three components:

- Setting goals;
- Letting work happen and comparing work completed against goals; and
- Deciding whether to change how the goals are being pursued."

In other words, there are three conditions that must be in place to say that performance is being managed: (1) clear, measurable goals; (2) measurement feedback provided to the performers in order to make decisions; and (3) the ability to control resources and conditions if the measurement feedback indi-cates need for a change.

In the field of human performance technology (HPT), this understanding of performance management provides a rationale for measurement and evaluation. We clearly identify the changes in performance we seek to produce. We measure and monitor the performance over time to determine whether our goals are being achieved, and at what rate. And we decide, based on the feedback provided by measurement, whether (and sometimes how) to change conditions when our goals are *not* being met.

This logic applies at two levels in our field. First, when we as performance improvement professionals are called in to help address a performance challenge, we must ensure that the three conditions described by Esque (2001) are in place. In fact, unlike many of our colleagues in the performance improvement field who conduct cause analyses at the front end of projects to determine what interventions to propose, Esque follows a simpler path: He asks whether the three conditions are in place. Then, because they usually are *not*, he helps clients to establish clear, measurable goals; continuous data-based feedback loops to the performers; and processes for making decisions to change when goals are not being achieved. Once these conditions are in place, he coaches performers and their management through a continuous, data-based performance improvement process.

At a second level, whether we choose to take such a "lean" approach to human performance improvement or follow a more traditional sequence starting with front-end analysis, the three conditions that Esque describes *should apply to our own performance as change agents*, as well as to the performance that we seek to improve. For us to be effective as performance improvement professionals, we need the feedback provided by measurement to determine whether to continue an intervention as planned—or to change. This is a simple cybernetic model of self-correction, inherent in both data-based performance improvement and in the fields of natural science and engineering upon which it has, at least in the past, been modeled. In the same way, this self-correcting approach is the raison d`etre for *evaluation*, the reason for its very existence, as described in this chapter.

A DIVERGENT PERSPECTIVE

This chapter takes a somewhat different approach to the discussion of performance measurement and evaluation, overlapping at points with some of the more conventional discussions provided in this volume, but also stepping outside of mainstream measurement and evaluation to highlight several key ideas. It covers much of the same ground as Binder's (2001) article on "a few important ideas," as well as elements of Binder's (2002–2004) online column on measurement and evaluation entitled: "Measurement Counts!"

While the field of HPT originally emerged from the natural science of behavior (Binder, 1995), with its focus on standard units of measurement and replicable descriptions of procedures similar to accepted practice in the physical sciences, HPT has come to encompass a wider array of technical and conceptual inputs, many from the so-called "softer" fields of education and the social sciences. These other fields have introduced methods and approaches to measurement and evaluation that do not always align with generally accepted criteria in the

natural sciences or engineering (Johnston & Pennypacker, 1993), especially with regard to the selection of measurement units and procedures. The principles and concepts presented in the current chapter reflect the author's background and perspective and are as much as possible grounded in the philosophy and practice of natural science. One of the advantages of sticking closely to principles of natural science is that, in many respects, we can demystify measurement and evaluation and make it more accessible to front-line performance improvement practitioners. While this might seem, at the outset, counter-intuitive, continue reading to discover whether or not you think it is a fair statement. In many respects, natural science approaches to measurement and evaluation are simpler in concept, and less encumbered by statistical models and theoretical baggage, than are many approaches derived from the social sciences.

TERMINOLOGY

We use many terms in the field of performance measurement and evaluation, some of which have been defined in detail and with great technical sophistication elsewhere in this volume. For purposes of this chapter, here is a short list of concepts, defined with the intention of eliminating confusion, and appealing as much as possible to plain English explanations.

Measurement

Measurement is the process by which we identify the dimension, quantity, or capacity [of a thing] (*American Heritage Dictionary*, 2006). In the field of performance improvement, this term refers to the identification of *what* to count (business results, work output, and/or behavior); selection of relevant quantitative units of measurement (such as simple counts, kilograms, meters, liters, or other measures); and collection of data expressed in those units. For example, we might identify *a successful business proposal* as a countable work output and include criteria that define a "successful" proposal. We can then count successful proposals over successive time intervals prior to intervention to determine "baseline" levels of productivity. We might additionally count *unsuccessful* proposals and use the "success ratio" of successful to unsuccessful ones as a secondary measure. Once we have decided on an intervention to improve business proposal productivity and quality, we can continue to count during successive time intervals to monitor whether or not the quantity of proposals and/or the ratio of successful to unsuccessful proposals is improving. This is not very different from keeping score in a sporting event, after first defining what constitutes a score, an error, a foul, and so on.

Evaluation

Evaluation is a process by which we evaluate or ascertain the value or worth of [a thing]. (*American Heritage Dictionary*, 2006). In performance improvement, we use measurement, plus some sort of evaluation design, to determine the impact and worth of an intervention. If, for example, measurement shows that the proportion of successful proposals as well as the total number of proposals completed per month accelerate after an intervention, and if we also measure the dollar value of successful proposals (and perhaps the average unit cost for submitting proposals), then we can determine (that is, *evaluate*) the worth of the intervention by calculating the increased number and proportion of successful proposals and the dollar value of the increase. This process would yield what is often referred to as an estimate of return on investment (ROI) if we were to compare measures before the intervention with measures following the intervention.

Performance Analysis

Analysis is another term used frequently in the literature of performance improvement and performance evaluation. The *American Heritage Dictionary* (2006) defines *analysis* as "the separation of . . . a whole into its constituent parts for individual study." In our field, the term analysis can mean many different things, depending on what is being analyzed. In this chapter, we will first discuss *performance analysis*, which breaks down human performance into its elements as a way of describing the performance we wish to improve and developing ideas about how we might improve it. Performance analysis forms a foundation for measurement strategies and tactics described later in this chapter.

Functional Analysis

A second type of analysis, called *functional analysis*, may be unfamiliar to some HPT practitioners, but derives from a basic principle of the natural science of behavior (Binder, 1995; Johnston & Pennypacker, 1993; Skinner, 1953). Functional analysis (or functional definition) uses measurement to determine what impact, or *function*, a variable (or behavior influence) has in relationship to performance, for example, the impact of providing job aids on the frequency of correctly diagnosing equipment failure. In the literature of behavior science, a "reward" arranged to follow a specific behavior can only be called (or said to function as) a *reinforcer* if data show that it results in an increase in the behavior it follows (Skinner, 1953). Similarly, in HPT our interventions can only be considered effective if we can demonstrate through measurement and evaluation their impact on performance. In other words, functional analysis is the *actual demonstration of function or effect*, using measurement and evaluation design,

rather than the assumption, perhaps based on prior research or experience, that a particular intervention "works."

While it might seem academic to introduce the term *functional analysis* in this context, there is an important reason for doing so. As managers or performance improvement specialists, we often try to create or apply *recipes*—standard procedures or interventions that, based on previous research or application, are "known" to be effective. If there is a short list of important take-aways from this chapter, it should include the recognition that *there are no sure-fire recipes*. We can never know in advance from scholarly research, or from prior real-world successes, whether or not a particular program, initiative, method, or intervention will work in the next case to which we apply it. We don't know whether a teaching program that worked with one group will be successful with all those in the next group. We don't know whether a feedback system that works in one setting will work in another, and so on. Individual and group differences, cultural variations, and many other factors often conspire to make ineffective, or to mask the effects of, procedures and programs that have previously proven successful. That is the most important reason for measurement and evaluation in HPT practice. We must continue to monitor and adjust our interventions, based on measurement feedback.

The best way that we can use prior experience and the findings of scholarly research is to formulate "good next bets" about what is likely to work in a given situation. We select programs and interventions based on scholarly research, prior experience in our own organizations, or best practice reports from others. But, as Esque's (2001) approach to performance improvement makes clear, we need to use the feedback provided by measurement to be sure what we are doing is effective here and now or to make decisions to change when the results are not as hoped. Functional analysis, according to which a program or variable can only be said to be effective when it is measurably shown to be so, is a core scientific principle that applies equally well to real-world performance improvement practice.

Research

A final term, *research*, deserves some discussion here, if only because there is frequent reference in the field of performance improvement to "research-based methods." A simple definition of research is "systematic investigation to establish facts" (Wordnet, 2006). As performance improvement specialists, we should make every effort to apply what is known from systematic and scholarly research to design our "best bet" interventions, based on the known "facts" about different types of programs and procedures. This is how as practitioners we can take advantage of formal research findings.

Hypothesis-Testing Research. Often scholarly research uses an hypothesis-testing approach in which conditions are arranged to test whether a particular

program, variable, or intervention has a specific, hypothesized effect. It is often possible to isolate and test the impact of elements that one might typically combine to form a single, complex intervention in the field. Basic scholarly research often uses statistical models, comparing *average* effects of different interventions, or evaluating the relative effects of variations of an intervention across groups or individuals. This approach is often neither practical nor particularly useful in applied settings, since our goal in most field applications is to improve the performance of *all* individuals or groups whose performance we are attempting to improve, not merely to demonstrate the relative effectiveness of different types of interventions under specific conditions. Nonetheless, scholarly hypothesis-testing research can still be helpful when we are attempting to assemble programs or interventions composed of multiple elements or variables. It can provide guidelines for what we might try in our "best bet" interventions and enable us to improve the likelihood that our initial designs will be effective.

Inductive Reasoning Research. Another type of research, more closely resembling and useful for practical application, is what would traditionally be called *inductive research*: the accumulation of multiple cases (individuals, groups, or others) in which changing a particular variable produces the desired results, to the point at which we feel confidently able to generalize from many successful "replications" to new but similar situations. When researchers (or practitioners) describe performance and its conditions clearly enough so that others can reliably repeat their procedures, and when they use standard units of measurement with clearly defined evaluation designs (Johnston & Pennypacker, 1993), it is possible to become more and more confident over time about what works in particular situations and about variations that might be most likely to succeed under different conditions. The idea is that we "induce" general rules or guidelines by accumulating multiple cases that resemble each other in critical features.

With this inductive approach in mind, practitioners should make every effort to carefully define performance and conditions and to use standard, repeatable measurement and evaluation procedures so that it becomes possible to generalize the results of one project or case to another and to accumulate cases over time to build confidence about the likely impact of specific types of programs or variables. Again, we can use such information to select "best bet" interventions or designs, and then make changes going forward as measured results provide feedback. Whether conscious or otherwise, this is what we all do as practitioners when we continue to refine our ability to predict what will work in different situations or with different types of performance. And as a field, to the extent we carefully describe performance, our procedures, and our measurement methods, we will be able to accumulate growing bodies of useful, prescriptive research. We'll have better and better ideas about "best bet" procedures and interventions to try.

A NATURAL SCIENCE FOUNDATION

Historical Roots

In this chapter and elsewhere (Binder, 1995), I have repeatedly referred to the "natural science" of behavior. By this, I mean the new science created by B.F. Skinner (Bjork, 1993) and his colleagues, in which the fundamental unit of measurement was rate of response (count/time), and the methodology used in the laboratory focused on the analysis, prediction, and control of behavior in the "individual organism" (Skinner, 1938). This science led to breathtaking discoveries and developments that included intermittent schedules of reinforcement, behavior shaping through reinforcing successive approximations to desired behavior, stimulus fading, programmed instruction, performance management, and the methods of behavior therapy. The International Association for Behavior Analysis is the growing and vital home for both basic researchers and field application of this science, and founders of the International Society for Performance Improvement (ISPI)—originally the National Society for Programmed Instruction—included many professionals who were applied behavior scientists in that tradition.

The reason for mentioning this aspect of our performance improvement lineage is to highlight the value of:

- Research and practice that employ standard and universal ("idemnotic") units of measurement rather than self-referencing ("vaganotic") indicators such as percentage correct or rating scale scores whose meanings can vary within or across applications (Johnston & Pennypacker, 1993),
- A focus on analysis and evaluation methods that reveal impact on individ-ual performance rather than averaging across groups (Binder, 1995), and
- Measurement as a continuous feedback loop, in contrast to one-time "validation" of methods (or recipes) and subsequent application without ongoing measurement feedback (Binder, 2001).

These are among the essential elements of HPT at its best, directly inherited from the science of behavior.

Role of Measurement

While earlier sections of this chapter have at points addressed the purpose or role of measurement in performance improvement, let us be very clear about the three typical purposes or types of measurement that we generally find in the development or management of human behavior.

- *Validation.* As suggested above, measurement often occurs in the context of research studies or best practices initiatives in which data collection and analysis serve the role of "validating" a particular program, type of

intervention, or variable's effect. While such work can, indeed, provide good input for designing "best bet" performance interventions, validation studies simply *cannot* guarantee that any particular program or type of intervention will apply in new situations, or even in very similar situations with what might seem to the casual observer to be "slight" variations in conditions or performers. For effective day-to-day management or performance improvement, we need to evaluate each time we intervene.

- *Accountability.* Much of the data collected in organizations and schools is intended to "hold people accountable"—whether the performers themselves, managers, or performance improvement specialists. Often such data are collected and stored in spreadsheets, databases, learning management systems, or other "containers" so that people can later retrieve the data "in case" they are needed. However, such data are not often collected or organized and stored in ways that can support frequent decisions about whether, when, or how to change conditions to improve performance. By the time we obtain test scores in most courses, it's too late to change procedures. By the time we look at spreadsheet summaries of "results" weeks or months after an initial intervention, it can be too late to change that intervention in a cost-effective way. While not all data collected for accountability are so difficult to use for other purposes, they do not often support the sort of agile decision-making and course correction that Esque's (2001) approach suggests.
- *Decision making.* Following the notion that measurement can and should provide information for a feedback loop, intended to support mid-course corrections and continuous improvements, the primary purpose of measurement and evaluation in performance improvement ought to be decision making. If this is true, then we should try to collect data frequently enough and display and use it in ways that allow us to adjust conditions and resources to optimize the pace, quantity, and ultimate impact of any performance change that occurs as a result of our programs or interventions. This is the same rationale as emerged from Skinner's (1938) science of behavior in which researchers continuously adjusted experimental conditions for individuals to maximize the pace and degree of behavior change in the desired direction.

In performance improvement we likewise want to be able to use measurement and evaluation to continue changing our programs and interventions until we "get it right."

Units of Measurement

Mention of *standard units of measurement* earlier in this chapter, and in previous publications (Binder, 2001), deserves expansion. In the natural

sciences, for scientists to be able to compare results from one experiment to another or to contribute to a coherent accumulation of scientific knowledge, there is an insistence on using standard, universal, and objective measurement dimensions and units. In fact, one could argue that many of the most important advances in science over the centuries have arisen from development of new, standardized measurement units and tools.

In the same way, if the field of performance improvement is to achieve the status of a true technology in the way that various fields of engineering have produced certifiable technologies, we must use standard dimensions and units of measurement. What this means (Johnston & Pennypacker, 1993) is that the measurement units and dimensions that we use to validate and make decisions about performance improvement programs and variables must mean the same from one situation to the next, from one individual to the next. Otherwise we can make nothing but very weak statements about the impact of our efforts.

Some measurement dimensions or units vary in their meaning from situation to situation. Good examples include percentage correct (Binder, 2004a) and average scores from Likert rating scales (Binder, 2003).

Percentage Correct. The trouble with percentage correct is that we cannot uniquely describe actual performance from the score. We don't know *how many opportunities* the performer was given to respond, *how many responses* were correct, or *how much time* it took to respond. Two people can achieve exactly the same percent correct scores on the same test, but with very different levels of performance, because percent correct ignores the time dimension. The same percent correct score can indicate quite different levels of performance from one situation to another, which is why accuracy-only tests, for example, are often very poor predictors of on-the-job performance. An additional, and often-confusing, aspect of percentages is that an increase by a given percentage (for example, adding 20 percent to 100, resulting in 120) is not equivalent to a decrease of the same value in percentage (for example, subtracting 20 percent from 120, resulting in 96).

Rating Scales. The results of evaluation using rating scales can also vary in meaning from one situation to another. First, the numbers on the scale are themselves relative, not absolute quantities. They simply indicate more than (>) or less than (<) lower or higher levels on the scale, respectively. In fact, the *numbers* on rating scales are actually *categories*, not quantities that can be added, multiplied, or otherwise manipulated with meaningful numerical results (Binder, 2003). Consequently, when they are combined into quantities and then averaged (for example, a score of 3.2 out of 5), the average numbers have no relationship to objective performance. It would be far more useful to use rating scales, if necessary, by counting and reporting the *numbers* of

people who assign each rating value, as in "thirty-two out of seventy people said service was excellent or good, while fifteen said it was average, and twenty-three said it was below average or poor." These numbers at least describe results in standard units—the actual counts of people. We can *directly* compare these counts, and their proportions, with other results quantified in the same way.

The general point here is that if we use standard measures (count, time, weight, volume, distance, and so forth), we will be able to evaluate results based on quantities that are standard across settings and applications—and that are therefore more likely to help us communicate and reliably contribute to accumulating knowledge of what works. As you will see in a following section, if we can describe business results, work outputs, and/or behavior using standard measurement dimensions, then we will be able to conduct measurement and evaluation in an objective, meaningful, and repeatable way, comparable to measurement used in natural science.

Key Concept: Calibration

Another concept from natural science that might be helpful for those attempting to measure performance and evaluate the impact of efforts to change it in the "real world" is *calibration*. Wikipedia defines calibration as "the process of establishing the relationship between a measuring device and the units of measure." In general, calibration is the stage in any measurement process whereby we check to be sure that the tool we are using accurately measures what we want it to measure. This concept can be applied at various levels in measurement and evaluation of performance improvement.

Validity of Analysis Unit. First, are we measuring what we intend to measure? Sometimes we're not certain whether we've chosen the right unit of analysis. For example, should we count lines of code written, or some other output, in order to measure the productivity of programmers? In other words, if we find ways of increasing lines of code, will we be contributing to the productivity of the programming team? (The generally accepted answer to that question is, "No." We need to identify some other unit to judge and count, since code efficiency is an important aspect of programming productivity not reflected when we count lines of code.)

Reliability of Data. Once we've chosen something to measure, are we reliably measuring it? For example, when measuring the behavior of people in a customer call center by observing and counting, we need to ascertain whether two or more observers are counting the same instances of behavior. We compare data collected by two observers and calculate inter-observer reliability (Daniels & Daniels, 2004, p. 143), the degree to which two people are observing and

counting the same behavior. This method calibrates the reliability of our instruments, in this case human observers with checklists and pencils. It's important to note, however, that two observers could be equally inaccurate but still agree in their measurement results. This would be a case of inaccurate but reliable (consistent) measurement.

Sensitivity of Procedures. Another aspect of calibration is related to the *sensitivity* of our measurement procedures. In the case of a microscope or a telescope, one might be able to observe more valuable or useful levels of detail at one magnification versus another. In some cases, the higher degree of magnification may actually be "too sensitive" for the purpose to which it is applied. We need to determine which level yields what type of information and which might be more useful for the purpose intended.

Similarly, if we are measuring human performance, the interval over which we count or the *counting period* (per hour, per day, per week, and so on) and the "chunk size" of what we are counting (for example, individual parts, subassemblies, or entire units) might make a difference as to what decisions we can make and how useful they might be. Among other things, the counting period determines how often we can make data-based decisions, since we need several data points in a row to determine the average level and the trend of the data. Similarly, when setting criteria for which work outputs are acceptable and which are not, it's important to determine which criteria will be more indicative of overall quality.

Refining the Measurement Plan. These are often decisions that, in the beginning, can be made only on the basis of pilot or trial runs or observations and analysis of collected data, both numerically and graphically, for the purpose of calibrating one's measurement procedures and tools. While calibration has been an important element of the quality management literature, it has not always been part of performance improvement practice. In general, it is important to recognize that metrics and measurement methods that you choose might need to be adjusted and refined during the early phases of any initiative or program evaluation process in order to be sure you are reliably measuring what you think you are measuring and that the data you collect are useful and help to inform good decisions, cost-effectively and practically. It might not always be clear in the beginning what to measure either. For this reason, it is often helpful to measure and graph results in more ways than you will after an initial calibration period, to determine what measures and presentations of the data turn out to be most indicative of what you are attempting to measure and most helpful for making decisions. These initial attempts and revisions of your measurement approach might require a number of adjustments, and it is good to plan for some time at the beginning of any project or effort for reviewing initial data,

summarizing and graphing the data in various ways, and possibly adding to or changing what and how you measure performance.

MEASUREMENT AND THE PERFORMANCE CHAIN

As discussed earlier, performance analysis is an essential prerequisite for performance improvement. We analyze performance by identifying the elements of what we call the *performance chain* (Binder, 2005). The performance chain shown in Figure 1.1 depicts how behavior produces organizational results and the behavior influences that make desired behavior likely to occur.

Figure 1.1 The Performance Chain.

We typically begin this analysis by identifying the individual or team *work outputs* that contribute to desired *organizational results* and then specifying the *behavior* required to produce those work outputs. The process of performance improvement is when we identify and attempt to design or manage the combination of *behavior influences* needed to establish, support, or accelerate desired behavior that will produce the work outputs that contribute to organizational results. These four elements comprise the performance chain. This is a simple model that multiplies many times in the context of real-world organizations and complex work processes. Many team and cross-functional processes are comprised of dozens or perhaps hundreds of these chains, linked end-to-end (where the output of one chain is the input to the next) or running in parallel. At whatever the organizational level, or however complex the performance we are seeking to improve, the elements of the performance chain give us two important types of linkage:

- *Behavior influences and behavior link to outputs and business results.* Work outputs describe "what the organization needs from its people" and provide the important linkage between the activity (behavior) of people and the results they need to achieve for their organizations. Once we understand what outputs are needed, we can discover or plan for behavior to produce them and then assemble the behavior influences (Binder, 1998) needed to establish, support, or accelerate that behavior. If our analysis of the linkage is accurate, we should be able to improve behavior to improve work outputs and thereby improve organizational results.

- *Units of analysis link to measurement.* The performance chain provides a convenient way to think about what we can measure. In the elements of behavior, work outputs, and organizational results, it points to units for analyzing performance that can be measured using the appropriate dimensions or units of measurement.

Organizational Results

Business executives and owners generally have ways of quantifying the organizational results they seek to achieve. Business experts and consultants sometimes help organizations determine what measures to use and at what level. For example, Kaplan and Norton's (1996, 2004) *balanced scorecard* methodology recommends cascading sets of measures from the top of the organization down through functions and departments to help define goals and monitor progress in strategic and tactical planning and execution. Others within our own field of HPT, most notably Rummler (Rummler & Brache, 1990; Rummler, 2004), have suggested systematic methods for assigning indicators and measures, most notably those that allow evaluation of cross-functional processes.

While not all measures of organizational results are equally sensitive, useful, or expressed in standard units, performance improvement professionals—depending on their roles and positions in the organization—are often given these metrics by organizational management as targets for improvement. Our jobs are often framed as doing something with the human resources of the organization to achieve or accelerate progress toward specified business results.

Work Outputs

Work outputs are (or should be) the focus of our measurement and improvement efforts. We are often asked to improve productivity in a department, to increase the efficiency and productivity of processes that incorporate many milestones or "sub-outputs" along the way, or to enable a new performance designed to produce certain outputs (for example, problems solved for users, signed contracts delivered by sales people). Because one of the most powerful contributions of HPT as a field has been the understanding that outputs (or *accomplishments*), not behavior, should be the focus of our efforts and starting points for our analyses (Binder, 2005; Gilbert, 1978, 2007), our challenge is to help define and measure valuable work outputs that contribute to organizational results and then work to improve their quality or quantity, timeliness, and so on.

Defining Criteria. When we measure outputs, we usually need to define criteria for *good ones*—which might specify qualitative dimensions or successful

outcomes that define them as acceptable. *Successful* sales presentations, for example, are those that lead to the next desired step in the sales process such as a customer request for a proposal. *Acceptable* responses to customer queries might be qualified as those that are timely, accurate, and result in the customer's saying that her problem has been solved. *Good* executive decisions of a particular kind might be those that are backed up by financial data, are sufficiently specific to be executable, and are linked to key performance indicators for the business. For any manufactured work output, quality and customer-acceptance criteria might apply, and so on.

Gilbert (1978, 2007) used the term *requirements* to describe what we are calling criteria, and categorized them into three sets of *quality*, *quantity*, and *cost*. To translate requirements more easily into measurement, Binder (2001) described "countable units corresponding to Gilbert's requirements" (p. 22) such as count of accurate and inaccurate items, count of timely or untimely outputs, or count of outputs meeting cost criteria.

Counting Output. The point is that once we have assigned criteria for judging a *good* output, we can count that output. While simple counting is not always the best way to measure work outputs, it is in many cases the simplest and most straightforward. We can monitor to see whether the counts per counting period of "good" ones go up and "bad" ones go down. In some cases (such as with resolved customer problems), we might want the total count per time interval to increase while the count of customers who say they are pleased by the service remains stable or increases. In some cases we are focused on units of volume or weight or we want timely delivery of process outputs that meet quality and cost criteria.

Behavior

Behavior is perhaps the most difficult and often the most expensive element of the performance chain to measure. We don't always need to measure behavior. If our intervention produces desired outputs at an acceptable level or accelerates outputs as planned, then we need not measure behavior. On the other hand, sometimes for diagnostic reasons or because we need to be sure that outputs are being delivered in the *right* way, we must measure behavior, if possible.

Automatic Measurement. Sometimes behavior can be measured automatically, which makes the measurement process both easier and less expensive. More and more automated systems exist for potentially capturing behavior measures, perhaps the most ubiquitous being the measurement of user and customer behavior on the Internet. Online systems can now count mouse-clicks, page visits, and other behavior of users in ways that allow web designers and business people to monitor the impact of changes in systems, content, or navigation on

websites. By automating the measurement of computer usage, we are actually turning behavior (key-presses or mouse-clicks) into simple outputs (switch closures) that we can count. But that's something of a technicality. Similarly, in security systems and other electronic environments that monitor door-openings, cardkey swipes, and other user activities, measurement is straightforward. Most sophisticated assembly lines have mechanisms that automatically turn behavior into countable mechanical or electronic events.

Observing Behavior. In many cases, especially those involving face-to-face interactions between humans, behavior is much harder to capture for measurement and evaluation. In those environments, such tools as behavioral checklists for observing or self-monitoring become necessary. Specialists in behavior management (Daniels & Daniels, 2004) have devised many procedures for judging and counting desired and inappropriate behavior. While some measures of behavior conform to criteria for standard and universal measurement units (those that always mean the same thing and can be compared across situations), others, particularly those involving rating scales or percentage calculations, fall short of natural science standards. In general, we encourage practitioners to identify criteria that distinguish between acceptable and unacceptable instances of behavior or among different classes of behavior, so that observers or self-observers can learn to reliably count instances and sum them over appropriate periods of time (per minute, per hour, per day, per week). Often we use behavioral checklists to tally behavior of different types. For some applications, carrying small notebooks for collecting tallies or using such devices as golf wrist-counters can make data capture easier.

Self-Monitoring. A type of behavior measure that generally escapes discussion among managers and performance improvement professionals is *self-monitoring* (Binder, 2004b). We can use self-monitoring to count behavior (or outputs) produced by the person counting his or her own performance. While one might doubt the reliability of one's counting one's own thoughts, feelings, or actions, research has demonstrated remarkable orderliness in self-monitoring, particularly if there is no incentive for the performer to "fake" the data. Sometimes on-the-job criterion-referenced learning programs or self-monitored fluency training (Binder & Sweeney, 2002) turn measurement procedures over to the learner, with dramatic results: by becoming responsive to their own measured performance, participants take enthusiastic control of their own learning processes, much like athletes monitor their own improvements in performance through continuous measurement and feedback.

Self-monitoring is often most powerful when managers or others, interested in improving their own behavior, count specific actions or activities throughout the day. The author, for example, has occasionally counted his own use of positive and

negative feedback delivered to staff as a means of monitoring efforts to use a more positive management style. When compared with other means of measuring the behavior of managers, supervisors, or others as they behave in relation to other people, behavior self-monitoring can be an attractive option.

BEHAVIOR INFLUENCES: THE INDEPENDENT VARIABLES

As the diagram of the performance chain in Figure 1.2 depicts, the factors that affect or influence behavior to produce work outputs and results are called *behavior influences*. These are the many different conditions in the performer's environment and techniques, tools, or methods that we can arrange to influence behavior. The list of such variables can be extremely long, especially if we take the relatively transient fads or "flavors of the month" in HRD or management development into account. How, in the end, we can make sense of these many different variables has been the focus of countless articles and models of performance improvement over the decades (Wilmoth, Prigmore, & Bray, 2002). This author uses the Six Boxes® Model (Binder 1998, 2005), a plain English framework that evolved from Gilbert's (1978) behavior engineering model (BEM).

Expectations and Feedback (1)	Tools and Resources (2)	Consequences and Incentives (3)
Skills and Knowledge (4)	Selection and Assignment *(Capacity)* (5)	Motives and Preferences *(Attitude)* (6)

Figure 1.2 The Six Boxes® Model.

To our knowledge, the Six Boxes Model is a *comprehensive* framework that encompasses every variable that can have an influence on behavior. Using this model, we describe all the elements in a performance improvement program or initiative, and categorize them into the six cells of the model. This is a convenient, powerful yet simple way to understand what a scientist would call the *independent variables* that we as managers or performance improvement

specialists configure to provide the most cost-effective impact on performance that we can arrange.

While a thorough description of the Six Boxes Approach (www.SixBoxes .com) is beyond the scope of this chapter, suffice it to say that we use it to better understand the programs we design and to better predict what changes we can make in the factors influencing performance that are likely to produce the desired outcome. When we evaluate programs or interventions, we are in effect evaluating the impact of performance systems comprised of variables that can be described and organized according to the logic of the Six Boxes Model. With experience or based on research, the model can often help us to determine "best bet" changes in performance programs (such as clearer expectations, better tools, rewards for doing the right thing, and so forth) likely to accelerate progress toward the desired outcome.

STANDARD DATA DISPLAY FOR DECISION MAKING

It should not shock any reader to know that anyone can "lie" with charts and graphs. Most professionals involved in measurement and evaluation of performance interventions summarize and present their data in graphic form. Some even use graphic displays to analyze and support *ongoing* decision-making about performance improvement. Some authors (Jones, 2000) have turned the phenomenon of distorting facts using graphic display into good humor; others have emphasized the positive potential of graphic display for highlighting important information or conclusions (Tufte, 2001). Those of us involved with making decisions about performance improvement interventions will benefit from keeping a few key distinctions in mind.

Stretch-to-Fill Graphs Versus Standard Graphic Display

For those accustomed to using PowerPoint or other software capable of creating graphs, the *stretch-to-fill* or *fill-the-frame* phenomenon is familiar. In fact, some even use it to advantage as a tool for persuasion about the size of effects and so forth. When we specify the ranges of the data we wish to graph and the type of graph we wish to use, the software generally selects the highest and lowest values on the scales to frame the data to fill the screen or a piece of paper. From one point of view, this is an effort to maximize visual attractiveness and best use of graphic *real estate*. However, because every graph created in this way consists of customized scales and distances between values, the proportions, angles, and distances equaling a given unit of measurement generally differ from one graph to the next. This means that the viewer must look carefully at the actual numbers on the graph to truly understand rates of change (trends), sizes of effects, ratios between sets of numbers, and so on. Visual comparisons between graphs

become impossible or deceptive, since the scales and proportions differ. While a standardized picture of data might, indeed, be worth a thousand words or more, idiosyncratic stretch-to-fill graphs can actually *inhibit* accurate communication of quantitative results.

In contrast, using standardized graphic displays offers the same power of comparison as does a standard twelve-inch ruler or any other tool designed to provide visual representation of quantities. By "standardized" we mean graphs in which the distance between numbers on scales for one graph is the same as for another with which we are likely to compare it. In the most general case, we might hope that an entire literature of, say, feedback effects, might use the same graphic displays. While this is perhaps unrealistic, the point is that standard graphic communication can significantly improve communication of quantitative results, comparison between cases, and so forth. One can directly compare effects, trends, proportions, and other dimensions of the data without having to look so carefully at each "customized" scale.

Lindsley (1999) illustrated this point with numerous examples of his standard celeration chart (Figure 1.3), a powerful visual tool for understanding and presenting ongoing measures of behavior, outputs, or organizational results.

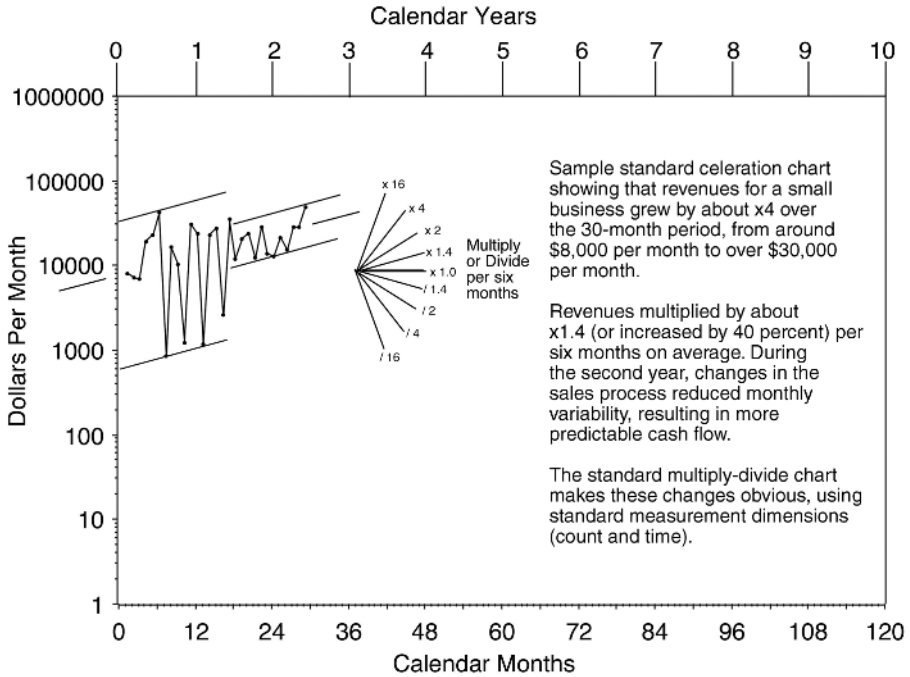

Figure 1.3 Standard Celeration Chart.

In his standard charting technology, Lindsley took advantage of the human ability to quickly scan and visually compare objects for similarities and differences. Without visual standards for the display of data, we place ourselves and others at a significant disadvantage and at risk of unintentionally misrepresenting or misunderstanding the results of our performance improvement efforts.

Uninterrupted Calendar Time Versus Sessions Scales

How often have you seen graphs on which the scale across the bottom was something like *sessions* or *observations*? Quick review of many journals and other publications displaying measures of human behavior or outputs over time reveals that they often ignore standard *calendar* time, substituting instead events displayed sequentially on the time scale, regardless of the varying time intervals between points. This means that if sessions or observations, for example, are sometimes scheduled daily and sometimes only every few days, we cannot tell from the display of the data because every data point simply appears on the next line in the sequence, not taking real-time distances between measures into account. In fact, the potential effects of missed days or sessions cannot be determined from such graphs, a phenomenon that can cause us to significantly misunderstand or even be oblivious to important time-related effects on performance.

If, instead, we use standard displays of calendar or clock time so that, when there is a day on which an event did not occur, we *skip* that line on the graph and go to the next (Binder, 2001), then we can see the impact of our interventions spread over a true representation of time. If a week of vacation intervenes, we can see any effect it might have had on performance after the week. If there were more than one session on a given day, we might see data displayed with two points on that day-line. In any case, we can see the effects in "real time" of our interventions, spaced as they are in actual time on our data display.

Equal-Interval Versus Multiply-Divide Graphic Scale

Many people recognize that certain quantities such as population tend to grow in multiples. For example, population of a given area, or of a given type of organism, is likely to multiply by a given factor (x2, x3, for example) for each successive period of time, rather than adding a fixed amount. This is why we have the "population explosion"—growth that is much more rapid than a fixed amount per unit time. Rather, it tends to be a fixed multiplication.

Lindsley and his colleagues (Lindsley, 1996; Pennypacker, Gutierrez, & Lindsley, 2003), based on research showing that human behavior also grows in multiples (or proportionally, a given percentage trend) rather than in additive increments, have perfected a graphic display (Figure 1.3) that takes advantage of this finding. They created the *standard celeration chart* over the course of more than forty years of research and development. The term "celeration" was coined

to reflect a standard measure of change, either AC-celeration or DE-celeration, quantified as a multiplicative or dividing trend per standard unit of time (per week, per month, per six months, and so on). A professional society, the Standard Celeration Society (www.celeration.org), exists for people who use this standardized graphic display in education, training, management, macroeconomic studies, and other fields.

While description of this chart might sound very esoteric, and perhaps only useful for the mathematically inclined, its design actually allows users as diverse as elementary school children and performance improvement specialists, performance coaches, and managers to make quick data-based decisions about trends and levels of measured performance. It is not necessary to know very much about the underpinnings of the chart to use it effectively.

What makes the standard chart helpful, with its multiplicative scale of counts up the left and its calendar time base across the bottom, is that any given angle on the chart represents the same rate of change, no matter what the level. And any given distance between two points on the chart reflects the same *ratio* (or multiplicative factor) between those two numbers, no matter what the levels. This means that one can learn to "read" the charts directly, without looking carefully at the numbers themselves, and rapidly understand the levels, bounce, or variability and trends in any data displayed on the chart. Such a graphic standard supports rapid display or sharing of data and rapid decisions. Its key features, with multiple examples, are presented in Binder's (2001) paper, available for downloading online.

SUMMARY AND CONCLUSION

While this chapter is by no means a complete discussion of performance measurement and evaluation, its intention has been to present the topic from a somewhat different perspective than usual, to introduce some new ideas, and to refer readers to additional resources for further study.

Key summary points to consider as you dig more deeply include the following:

- The most practical and directly useful purpose of performance measurement is to *make decisions* about whether or not efforts to improve performance are having the desired impact and whether or not to make changes before too much time has elapsed.
- Measurement provides *feedback* to performers, managers, and performance improvement specialists so that they can adjust their behavior and their efforts to improve.
- Being careful to describe our procedures and methods clearly and thoroughly enough *so that others can replicate them* will advance both

HPT practice and its scientific foundations in the most reliable and sustainable way.

- The *performance chain*, linking organizational results to work outputs to behavior and its influences, provides a good reference for what we might choose to measure (behavior, work outputs, and/or organizational results).
- Using standard measurement units rather than quantities with no reliable real-world reference (such as averaged rating scales or percentage correct) allows us to bring rigor and objectivity equivalent to that of natural science to our measurement and evaluation of performance.
- How we display our data using standard graphic presentations is as important as the data themselves. We should be careful and self-critical as we attempt to truly understand results from the graphs and charts we use to analyze and display them. The standard celeration chart is a powerful tool for graphic display of performance data.

Following these guidelines will enable practitioners and researchers alike to contribute to a strong foundation in practical performance measurement and evaluation and to the accumulated knowledge base of our field.

References

The American Heritage dictionary of the English language (4th ed.). (2006). New York: Houghton Mifflin.

Binder, C. (1995). Promoting HPT innovations: A return to our natural science roots. *Performance Improvement Quarterly*, *8*(2), 95–113.

Binder C. (1998). The six boxes: A descendent of Gilbert's behavior engineering model. *Performance Improvement*, *37*(6), 48–52.

Binder, C. (2001). Measurement: A few important ideas. *Performance Improvement*, *40*(3), 20–28. Available online at www.binder-riha.com/measurement_ideas.pdf.

Binder, C. (2002, March/2004, November). Measurement counts! www.performance xpress.org.

Binder, C. (2003, March). Using surveys and questionnaires. In *Measurement counts!* Online monthly column at www.performancexpress.org/0303/mainframe0303. html#title5.

Binder, C. (2004a, September). The dangers of percent: An example. In *Measurement counts!* Online monthly column at www.performancexpress.org/0409/mainframe 0409.html#titlemeasure.

Binder, C. (2004b, January). Counting one's own behavior and accomplishments. In *Measurement counts!* Online monthly column at www.performancexpress.org/0401/ mainframe0401.html#titlemeasure.

Binder, C. (2005). *What's so new about the six boxes?* A white paper available at www .SixBoxes.com/resources.html.Bainbridge Island, WA: Binder Riha Associates.

Binder, C., & Sweeney, L. (2002, February). Building fluent performance in a customer call center. *Performance Improvement*, *41*(2), 29–37.

Bjork, D. W. (1993). *B. F. Skinner: A life*. New York: HarperCollins.

Daniels, A. C., & Daniels, J. E. (2004). *Performance management: Changing behavior that drives organizational effectiveness* (4th rev. ed.). Atlanta, GA: Performance Management Publications.

Esque, T. J. (2001). *Making an impact: Building a top-performing organization from the bottom up*. Atlanta, GA: CEP Press.

Gilbert, T. F. (1978). *Human competence: Engineering worthy performance*. New York: McGraw-Hill.

Gilbert, T. F. (2007). *Human competence: Engineering worthy performance*. San Francisco: Pfeiffer.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Jones, G. E. (2000). *How to lie with charts*. Bloomington, IN: Indiana University Press.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard*. Boston: Harvard Business School Press.

Kaplan, R. S., & Norton, D. P. (2004). *Strategy maps: Converting intangible assets into tangible outcomes*. Boston: Harvard Business School Press.

Lindsley, O. R. (1996). Performance is easy to monitor and hard to measure. In R. A. Kaufman, S. Thiagarajan, & P. MacGillis (Eds.), *The guidebook for performance improvement: Working with individuals and organizations* (pp. 519–559). San Francisco: Pfeiffer.

Lindsley, O. R. (1999). From training evaluation to performance tracking. In H. Stolovitch & E. Keeps (Eds.). *The handbook of human performance technology* (2nd ed.). (pp. 210–236). San Francisco: Pfeiffer.

Pennypacker, H. S., Gutierrez, A., Jr., & Lindsley, O. R. (2003). *Handbook of the standard celeration chart*. Concord, MA: Cambridge Center for Behavioral Studies.

Rummler, G. A. (2004). *Serious performance consulting*. Silver Spring, MD: International Society for Performance Improvement.

Rummler, G. A., & Brache, A. P. (1990). *Improving performance: How to manage the white space on the organization chart*. San Francisco: Jossey-Bass.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Wilmoth, F. S., Prigmore, C., & Bray, M. (2002). HPT models: An overview of the major models in the field. *Performance Improvement*, *41*(8), 16–25.

Wordnet 3.0 (2006). Princeton University.