Title
Attributes of Good Measures

Permalink
https://escholarship.org/uc/item/5k14q7v0

Author
Stevens, D.F.

Publication Date
1990-12-01

LBL-30001

Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

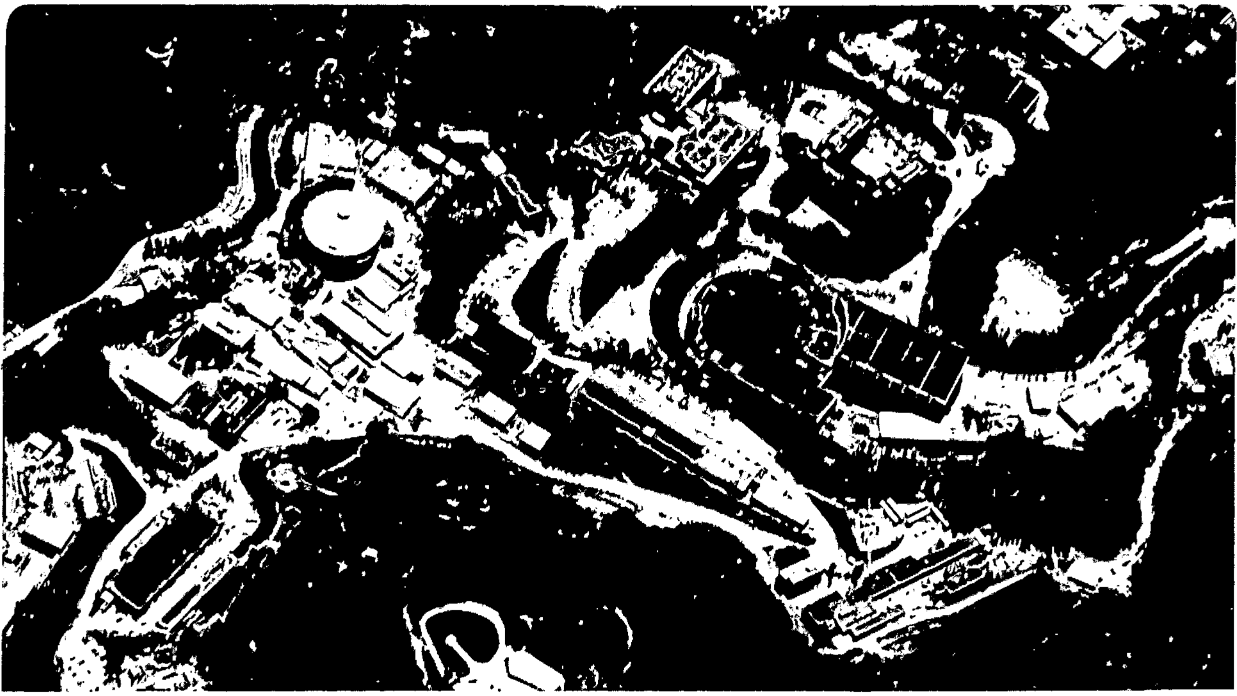
Information and Computing Sciences Division

To be presented at the 12th Annual IS
Performance/ Capacity Management Conference,
Mesa, AZ, March 4–8, 1991, and to be
published in the Proceedings

Attributes of Good Measures

D.F. Stevens

December 1990



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-30001

Attributes of Good Measures*

David F. Stevens
Information and Computing Sciences Division
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, California 94720

December, 1990

Prepared for the Twelfth Annual
IS Performance/Capacity Management Conference
Mesa, Arizona, March 4-8, 1991

* This work was supported by the U. S. Department of Energy under contract No. DE-AC03-76SF00098.

Attributes of Good Measures[1]

David F. Stevens
Lawrence Berkeley Laboratory
Berkeley, California 94720

> *When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.*
>
> Lord Kelvin, 1883

Introduction: Beyond Kelvin

The urge to measure is for some of us a fundamental need. We have been collecting numerical information ever since we were kids counting the pickets in the neighbor's fence or the number of steps on the way to school. We counted what there was to count, and left the interpretation, if any, to others. As adults, we have become Capacity Planners, spending much of our time recording and analyzing measurements and converting them into multicolored charts for presentation to upper management. (We even read newsletters and go to conferences to learn how to do it better.) In performing our chosen work we measure what we are speaking about, and express it in numbers—or so we think—and we are tempted to conclude therefore that our knowledge is neither meager nor unsatisfactory. But are we really measuring *what we are speaking about*? Or do we count EXCP's and speak of *input and output*? Measure clocktime and speak of *availability*? Measure service levels and speak of *saturation*? Count phone calls and speak of *user satisfaction*?

Before we leap to unwarranted Kelvinistic conclusions we might wish to consider what two other writers on the subject of knowledge have to say:

> *When you know a thing, to recognize that you know it, and when you do not know a thing, to recognize that you don't know it: That is knowledge.*
>
> Confucius, *ca* 500 BC

> *Where is the wisdom we have lost in knowledge?*
> *Where is the knowledge we have lost in information?*
>
> T. S. Eliot, *ca* 1930

Our profession provides us with such a profusion of measure-based numeric information that our interpretive processes are subject to overload. In our eagerness to satisfy Kelvin's criterion we assume our measures are meaningful, and that they measure what we expect and intend them to measure. We ignore Eliot's questions and we fail the Confucian test. In so doing, we often allow the measure to supplant the underlying reality, so that the object of our activity becomes improvement

[1] This work was supported by the U. S. Department of Energy under contract No. DE-AC03-76SF00098.

of the measure rather than improvement of the system it represents. To prevent this perversion of the process we must retain an awareness of the purposes for which we take the measurements and we need to know how well our measures meet those purposes: However "good" a measure may be by other criteria, it is not a Good Measure if it does not serve its purpose.

In an earlier paper[2] I postulated three fundamental reasons to measure: to support value judgements, to define models, and to detect, demonstrate, or monitor change. Since capacity planning deals primarily with the second and third of these, the attributes of interest here are those that support these purposes. For example, one element of a good capacity planning program is an early-warning system to give notice of impending difficulties. Measures employed for this purpose have less need to be *precise* than to be *timely, reliable,* and *of sufficient magnitude to attract our attention.* Attributes useful in other contexts may have little meaning for capacity management. Measures used to differentiate between nearly identical entities need to be extremely precise, for example, while those used in the development of approximate models can be correspondingly rough. Similarly, some measures need to be exactly right all of the time; others need only be approximately right most of the time. The appropriate precision for measurements of time will vary widely depending upon the speed of the reactions being measured. In the capacity management case, there are useful measures in terms of all intervals from nanosecond to year; it should be clear that equal precision is not desirable for all of them.

With this diversity in mind, we can begin our look at the generic qualities of good measures.

Of Goodness in Measures

From the brief discussion above, we can see that three qualities shared by good measures are *utility, trustworthiness,* and *timeliness;* two others are *simplicity* and *directness.* We shall consider all of these in a bit more detail. But first, a couple of cautionary notes about three *non*-attributes of good measures: precision, popularity, and accessibility.

We begin with *precision* because it is reasonably clear that the right amount of precision is a Good Thing. *More* precision, however, is not necessarily better. This is especially true when taking measurements that will be used in extensive calculations, such as when running a projection model; if the precision is greater than the accuracy, carrying it into extensive calculations can reduce the accuracy of the results, eventually to zero. Excess precision should also be avoided in reporting stand-alone measurements. It gives a false impression of detail, and obscures the fact that the low-order digits are often only an artifact of digital approximation to true values, and have negative significance, if any.

Popularity and *accessibility* are like precision in that they are not intrinsically bad, but they are in general no reason to choose one measure over another. In fact, the cynical might say that the popular ones are popular because they are accessible, and the accessible ones are accessible because they are the measures the vendor wants you to use. There is nothing wrong with following vendors'

[2] *New System Measures*, EDP Performance Review, January, 1985

implicit advice as long as you remember that the business of vendors is vending. However, these measures can be extraordinarily misleading—witness the continuing furore over lines-of-code as a measure of programmer productivity—as well as perniciously persistent, as is demonstrated by our inability to eliminate MIPS-based comparisons.

Usefulness

There is little point to taking measurements that don't contribute to our understanding of the state (or trend) of the system. To test the usefulness of a measure, we should ask questions such as the following:

- Does it tell us what's happening?
- Does it tell us what's wrong? what's right?
- Does it tell us what we want to know?
- Does it tell us what we need to know?

The answer to the first question is almost always "Yes", in an extremely limited literal sense; the answers to the other questions are more difficult to assess. I am reminded of an incident from the Good Old Days of computer performance measurement, when I was in charge of operating system performance for a reasonably large system. In the course of "improving" the system I introduced a change into a disk driver that caused CPU utilization to drop by 10%. According to the conventional wisdom of the time, that made it a Bad Change, because we all knew that Performance *was* CPU Utilization. In fact, however, it turned out to be a very Good Change that increased throughput significantly, for the disk driver had been doing some incredibly stupid wheel-spinning. The fault was not in the measure, for it did in fact measure the reality of CPU utilization, but in our simple-minded assumption that "high" was "good", ignoring the fact that effort is not necessarily synonymous with progress.

A measurement gives us the instantaneous value of the quantity or attribute being measured at the time the measurement is taken, but that may or may not be "what's happening" in any meaningful sense. There are two familiar caveats in the measurement business that underscore the distinction between measurement and reality, but we tend to forget them in our drive to boil our information down into digestible, one-page executive summaries. We have just been reminded of one of them—*viz.*, measurement of effort is not necessarily measurement of progress. The second is a generalization of the first: measurements don't tell you how *well* you're doing, only how *much*. (We have all seen a football team score 40 points (ordinarily deemed a successful effort) and lose, and we have all seen projects on which no work has been done listed as 90% complete on a PERT chart just because 90% of the scheduled time had elapsed.) In particular, "what's happening" to measured quantities may provide little or no insight into the other questions that address utility. The answers to these are more often "No" than many of us realize, for a measure is merely a number unless we know the context and the relationships between our measures and the realities of the system. In the situation mentioned above, the CPU measure told me what I (thought I) wanted to know, namely, that utilization was very high, but it failed to tell me what I needed to know, namely, that much of it was going to waste.

Good measures reflect reality in some known manner, so that we know, with respect to each measure, whether the big values, the little values, or the middle-sized values are the ones to be desired, we know how big "big" is, and we know whether the desirable trend is upward, downward, cyclic, or steady. But even good measures should not be considered in isolation. Unless we know what the measure is *really* measuring we cannot evaluate it properly. It is very easy to allow conventional "wisdom" to become a sort of tunnel vision in which the optimization of a measurement is pursued to the detriment of the underlying system.

Trustworthiness

A trustworthy measure is one that does not provide false indications. Elements of trustworthiness include repeatability, consistency, and (suitable) sensitivity.

Repeatability and consistency address the problem of whether a given reading always has the same meaning; i.e., whether the same fundamental conditions always give rise to the same observed values, and conversely, whether the same observed values always denote the same fundamental conditions. This is easier to achieve with objective measures than with subjective ones, for the latter are far more influenced by context. (If you doubt this, try the classic experiment of putting your left hand into a bucket of ice water and your right hand into a bucket of very hot water for five minutes. Then plunge both hands simultaneously into a third bucket, filled with lukewarm water. Your left hand will tell you the water is hot and your right hand will tell you it's cold.) A factor that can contribute mightily to consistency is the simplicity of the measurement procedure; a straightforward one-step process is far more likely to yield reproducible results than an intricate procedure with many decision points (ask any first-year chemistry student).

Good measures are also consistent in another sense: The set of "good" values remains the same, or at worst drifts predictably.

Sensitivity is like precision in that too much is as bad as too little. For a measure to be suitably sensitive, small changes in the observed values should correspond to small changes in the underlying fundamental conditions, and vice versa, and it should exhibit no singularities over its range of application.

Timeliness

A desirable attribute for those measures that indicate the health of a system—rather than merely giving its dimensions—is that they provide advance warning. In the parlance of the financial sections of the newspapers, *leading indicators* are far more useful than *lagging indicators.* The latter are necessary to convince some theoreticians that what they saw really happened, but do not help a pragmatist to prepare for disaster. If you would rather avoid a catastrophe than confirm it, you must have reliable leading indicators. (Although as Cassandra found out, even certain knowledge is not always good enough in the face of determined optimism.)

Simplicity

There are two ways in which simplicity can contribute to the goodness of a measure. The first, as noted in the discussion of trustworthiness, is in simplicity of the measurement procedure. The simpler the procedure, the more likely that we will get it right and that we will obtain the correct reading of the desired measure. Furthermore, no matter how good a measure is in other respects, if it is *too* difficult to take successfully, it will be abandoned. The second is simplicity of interpretation. Thus, for example, an absolute measure, where *more* is always better (or always worse) than *less*, is easier to deal with and interpret than a relative measure, where the definition of "good" depends upon what's happening in other places. (Perhaps that's why match play in golf has given way almost completely to medal play.) If the interpretation of a measure is rife with uncertainties, or easily capable of misconstruction, the measure will fall into disuse. (I make this claim with some trepidation, knowing that the practice of economic forecasting provides numerous counterexamples—of persistence, if not of success.)

Directness

Most of the traditional computer performance measures do not measure the actual events and conditions of interest; instead, they measure presumed *effects* or *causes* of those events and conditions. Thus we measure *queue length,* an effect of service processing, instead of directly measuring the server process[3]. Similarly, we use the *number of reruns* as a measure of the quality of the operational procedure, and *complaint count,* as a negative measure of the quality of a service activity. Commonly used cause measures are *multiprogramming factor* and *channel overlap,* as measures of productivity, and MIPS and MegaFLOPS as measures of raw power.

It is possible for one measure to fit more than one of these categories, depending upon what the investigator is interested in. Almost any of the resource utilization measures can be interpreted as a cause (with respect to saturation, and hence to long turnaround or response times), as an effect (of allocation and scheduling policies), or even as a direct measure (of volume).

Other measures are even less direct than effect and cause; they are *concomitants* of events or conditions of interest, that is, events or conditions not intrinsically of interest themselves, but that tend to occur in conjunction with, or at the same time as, events or conditions of interest. EXCP counts are such measures; they are system artifacts that occur with I/O activity, and are often used in chargeback algorithms as a substitute for a true measure of I/O activity, even though the actual relationship between EXCP's and bits moved is rather imprecise.

The usefulness of these measures diminishes with the length of the causal chain, sometimes to the point of demanding a considerable leap of faith or deductive logic before we can comfortably use

[3] Or of the quality of the scheduling process: the distinction is not always clear.

them. Even should a precise and calculable relationship exist between suggested cause and presumed effect, its demonstration can be extremely difficult to establish.

Thus, regardless of purpose and context, the best measures are those that measure directly the quantity or quality of interest. Direct measures are easier to use and are less likely to lead to misunderstanding and error than are indirect measures. Indirect measures may have sound theoretical foundations, yet be dangerously inaccurate in practice. An example from the history of aviation is the pressure-sensitive altimeter. The relationship between pressure and altitude under laboratory conditions is well understood, but in practice it fluctuates with the weather; and even in the theoretical case it provides only an *absolute* altitude (height above sea level) rather than the height separating the instrument (or the airplane bearing it) from the ground. Safe traverse over mountainous territory thus demands an accurate knowledge of the terrain over which the airplane is flying. Newer, radar-based equipment tells the pilot directly how far he is from the ground, independently of the barometric pressure or how far off course he may have drifted. Similarly, the *airspeed indicator* does a good job of informing the pilot how fast he is moving through the air, but provides only an approximate indication of how fast he is moving over the ground. (Stories are even told of light planes flying flat out but, because of the wind, actually moving backwards with respect to the ground!) Unfortunately, relatively few of the common computer performance measures are direct. Those that could be, such as turnaround time and availability, are frequently degraded into indirect measures by defining them from the system or data center point of view instead of from the user's point of view (but that's another whole issue).

Afterword

There are few really good measures available to the computer performance and capacity manager. Our measures are occasionally timely and trustworthy, but generally indirect (sometimes extremely so), often complex, and useful only to the extent we are aware of these other problems. But until we develop better measures, we need to use them as best we can. We must remember that the fundamental purpose of performance and capacity management is not the acquisition of information about our microenvironment, but the development of a level of understanding that will facilitiate the timely deployment of necessary resources. If, in our measurement programs, we keep the cautionary advice of Eliot and Confucius in mind—if we increase our awareness of what we really know about the measures and do not mistake information for knowledge—if we stop speaking about what we wish we were measuring and instead start to speak about what we actually measure—then we may, indeed, be said to approach the stage of science.