29
You Get What You Measure

You may think the title means if you measure accurately you will get an accurate measurement, and if not then not; but it refers to a much more subtle thing—the way you choose to measure things controls to a large extent what happens. I repeat the story Eddington told about the fishermen who went fishing with a net. They examined the size of the fish they caught and concluded there was a minimum size to the fish in the sea. The instrument you use clearly affects what you see.

The current popular example of this effect is the use of the bottom line of the profit and loss statement every quarter to estimate how well a company is doing, which produces a company interested mainly in short term profits and has little regard to long term profits.

If in a rating system every one starts out at 95% then there is clearly little a person can do to raise their rating but much which will lower the rating; hence the obvious strategy of the personnel is to play things safe, and thus eventually rise to the top. At the higher levels, much as you might want to promote for risk taking, the class of people from whom you may select them is mainly conservative!

The rating system in its earlier stages may tend to remove exactly those you want at a later stage.

Were you to start with a rating system in which the average person rates around 50% then it would be more balanced; and if you wanted to emphasize risk taking then you might start at the initial rating of 20% or less, thus encouraging people to try to increase their ratings by taking chances since there would be so little to lose if they failed and so much to gain if they succeeded. For risk taking in an organization you must encourage a reasonable degree of risk taking at the early stages, together with promotion, so finally some risk takers can emerge at the top.

Of the things you can choose to measure some are hard, firm measurements, such as height and weight, while some are soft such as social attitudes. There is always a tendency to grab the hard, firm measurement, though it may be quite irrelevant as compared to the soft one which in the long run may be much more relevant to your goals. *Accuracy* of measurement tends to get confused with *relevance* of measurement, much more than most people believe. That a measurement is accurate, reproducible, and easy to make does not mean it should be done, instead a much poorer one which is more closely related to your goals may be much preferable. For example, in school it is easy to measure training and hard to measure education, and hence you tend to see on final exams an emphasis on the training part and a great neglect of the education part.

Let me turn to another effect of a measurement system, and illustrate it by the definition and use of IQs. What is done is a plausible list of questions, plausible from past experience, is made, and then tried out on a small sample of people. Those questions which show an internal correlation with others are kept and those which do not correlate well are dropped. Next, the revised test is calibrated by using it on a much larger sample.

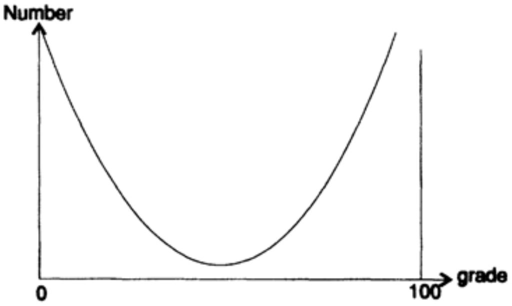

Figure 29.I

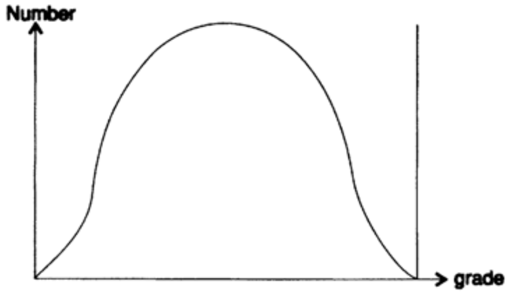

Figure 29.II

How? Simply by taking the accumulated scores (the number of people's scores which are below the given amount) and plotting these revised numbers on probability paper—meaning the cumulative probabilities of a normal distribution are the horizontal lines. Next the points where the cumulative actual scores fall at given percentage points are related, via a calibration table, to the corresponding points on the cumulative normal probability curve. As a result it is observed intelligence has a normal distribution in the population! Of course it has, it was made to be that way! Furthermore, they have defined intelligence to be what is measured by the calibrated exam, and if that is the definition of intelligence then of course intelligence is normally distributed. But if you think maybe intelligence is not exactly what the calibrated exam measures, then you are entitled to doubt intelligence is normally distributed in the population. Again, you get what was measured, and the normal distribution announced is an artifact of the method of measurement and hardly relates to reality.

In giving a final exam in a course, say in the calculus, I can get almost any distribution of grades I want. If I could make up an exam which was uniformly hard, then each student would tend either to get all the answers right or all wrong. Hence I will get a distribution of grades which peaks up at both ends, Figure 29.I. If, on the contrary, I asked a few easy questions, many moderately hard, and a few very hard ones, I would get the typical normal distribution; a few at each end and most of the grades in the middle, Figure 29.II. It should be obvious if I know the class then I can get almost any distribution I want. Usually, at the final exam time I am most worried about the pass-fail dividing point, and design the exam so I will have little doubt as to how to act, as well as have the hard evidence in case of a complaint.

Still another aspect of a rating system is its *dynamic range*. Suppose you are given a scale of 1 to 10, with 5 being the average. Most people will give ratings of 4, 5, and 6, and seldom venture, if ever, to the extremes of 1 and 9. If you give a 6 to what you like, but I use the entire dynamic range and assign a 2 to what I do not like, then the effect of the two of us is while we may differ equally in our opinion, the sum of the ratings will be 6+2=8, and the average will be 4—the effect of my opinion more than wipes out yours! In using a rating scheme you should try to use the entire dynamic range, and if you do you will have a much larger effect on the final average—provided it is done, as most such cases are, by blind averaging of the ratings assigned. Remember Coding Theory says the entropy (the average surprise) is maximum when the distribution is uniform. You have the most information when all the grades are used equally, as you may recall from Chapter 13 on Information Theory.

If you regard giving grades in a course as a *communication channel* then, as just noted, the equally frequency use of *all* the grades will communicate the maximum amount of information—whilst the typical use in Graduate Schools of mainly the two highest grades, A and B, greatly reduces the amount of information sent. I understand the Naval Academy uses rank in class, and in some sense this is the only defense against "grade inflation" and the failure to use the whole dynamic range of the scale uniformly, thus communicating the maximum amount of information, given a fixed alphabet for grades. The main fault with using rank as the grade is by chance there may be all very good people in a particular class, but some one of them will have to be at the bottom!

There is also the matter of how you initially attract people to the field. It is easy to see in psychology the people who enter the field are mixed up in their heads more than the average professor and average student in a college—it is not so much the courses do this, though I suspect they help to mix the student up further, but the initial selection does it. Similarly, the hard and soft sciences have their attractions and repulsions based on *initially perceived features* of the fields, and not necessarily on the actual features of the field. Thus people tend to go into the fields which will favor their peculiarities, as they sense them, and then once in the field these features are often further strengthened. Result—poorly balanced, but highly specialized, people—which may often be necessary to succeed in the present situation.

In Mathematics, and in Computer Science, a similar effect of initial selection happens. In the earlier stages of Mathematics up through the Calculus, as well as in Computer Science, grades are closely related to the ability to carry out a lot of details with high reliability. But later, especially in Mathematics, the qualities needed to succeed change and it becomes more proving theorems, patterns of reasoning, and the ability to conjecture new results, new theorems, and new definitions which matter. Still later it is the ability to see the whole of a field as a whole, and not as a lot of fragments. But the grading process has earlier, to a great extent, removed many of those you might want, and indeed are needed at the later stage! It is very similar in Computer Science where the ability to cope with the mass of programming details favors one kind of mind, one which is often negatively correlated with seeing the bigger picture.

The personnel employment department also has an effect on who is recruited into the system. If there is recruiting for research then the typical member of the personnel department in a big organization is not likely to want the right people. Good researchers, because the criterion is they have originality in Science and Engineering, also means typically they are original in other aspects of their behavior and dress—meaning they do not appeal to the typical recruiter from the personnel department. Hence, as at Bell Telephone Laboratories, usually the research people go out to do the hiring for the research area, and the personnel department shudders! This is not a trivial point, the recruiting of one generation determines the organization's next generation.

There is also the vicious feature of promotion in most systems. At the higher levels the current members choose the next generation—and they *tend strongly* to select people like themselves— people with whom they will feel comfortable. The Board of Directors of a company has a strong control of the officers and next Board members who are put up for election (the results of which is often more or less automatic). You tend to get inbreeding—but also you tend to get an organization personality. Hence the all too common method of promotion by self-selection at the higher levels of an organization has both good and bad features. This is still on the topic you get what you measure as there is a definite matter of evaluation, and the criteria used, though unconscious, are still there.

In the distant past to combat this inbreeding most Mathematics Departments (a topic I am more familiar with than for other Departments) had a general rule they did not employ their own graduates. The rule is not now widely applied so far as I can see—quite the contrary, there seem to be a tendency to hire their own graduates over outsiders. There have been several occasions when Economics Departments were so inbred the top management of the University had to step in and do the hiring over the professor's dead bodies as it were, in order to gain a reasonable balance in the University of differing opinions. The same has happened in Psychology Departments, Law, and no doubt in others.

As just mentioned, a rating system which allows those who are in to select the next generation has both good and bad features, and needs to be watched closely for too much inbreeding. Some inbreeding means a common point of view and more harmonious operation from day to day, but also it will probably not have great innovations in the future. I suspect in the future, where I believe change will be the normal state of things, this will become a more serious matter than it has been in the past—and it has definitely been a problem in the past!

I trust you realize I am not trying to be too censorious about things, rather I am trying to illustrate the topic of this chapter— you get what you measure. This is seldom thought about by people setting up a rating, measuring, or other schemes of recording things, and yet in the long run it has enormous effects on the entire system—usually in directions in which they never thought about at all!

Although measuring is clearly bad when done poorly, there is no escape from making measurements, rating things, people, etc. Only one person can be the head of an organization at one time, and in the selection there has to be a reduction to an simple scale of rating so a comparison can be made. Never mind humans are at least as complex as vectors, and probably even more complex than matrices or tensors of numbers; the complex human, plus the effect of the environment they operate in, must somehow be reduced to a simple measure which makes an ordered array of choices. This may be done internally in the mind, without benefit of conscious thinking, but it must be done whether you believe in rating people or not-there is no escape in any society in which there are differences in rank, power to manage, or what ever other feature you wish. Even on a program of entertainment, there has to be a first and a last performer—all cannot be equally placed. You may hate to rate people, as I do, but it must be done regularly in our society, and in any society which is not exactly equal at all points this must happen very often. You may as well realize this and learn to do the job more effectively than most people do—they simply make a choice and go on, rather than give the whole process a good deal of careful thought, as well as watching others doing it and learning from them.

By now you see, I hope, how the various scales of measurement effect what happens. They are fundamental yet they normally receive very little attention. To strengthen what I have been saying, I will simply tell you more examples of how the measurement scale affects the system.

Earthquakes are almost always measured in the Richter scale, which effectively uses the log of the estimated amount of energy in the earthquake. I am not saying this is the wrong measuring scale, but its effect is you have few really large earthquakes, 7 and 8, and lots of small ones, 1 and 2. Think about it. I do not know the distribution on Mother Nature’s scale, but I doubt She uses the Richter scale. Linear transformations, as from feet to meters, are not serious, but nonlinear scale transformations are another matter. Most of the time we measure stimuli applied to humans on a log scale, but for weight and height we use linear scales. Linear ones allow additivity easily, but for nonlinear scales you do not have this. For example, in measuring the size of a herd you are apt to count the number of animals in the herd. Thus you have additivity—adding two herds together gives the proper amount of the combined herd. If you have a herd of 3 and add 3 that is one thing, but if you have a herd of 1000 and add 3 it is quite another thing— hence the additivity of the number in the herd is not always the proper measure to use. In this case the *percentage change* might be more informative.

How, then, do you decide which scale to use in measuring things? I have no easy answer. Indeed, I have the awful observation while one scale of measurement is suitable for one kind of conclusion in a field, another scale of measurement may be more appropriate for some other kind of decision in exactly the same field! But how seldom is this recognized and used! Of course you may observe sometimes we quietly make a transformation when we apply a given formula, but which scale of measurement to use is a difficult thing to decide in any particular case. Much depends on the field and the existing theories, as well as the new theories you hope to find! All of which is not much help to you in any particular situation.

There is another matter I mentioned in an earlier chapter, and must now come back to. It is the rapidity with which the people respond to changes in a rating system. I told you how there was a constant battle between me and the users of the computer, me trying to optimize the performance for the system *as a whole,* and they trying to optimize *their own use*. Any change in the rating system you think will improve the system performance as a whole is apt to not work out well unless you have thought through the response of the individuals to the change—they will certainly change their behavior. You have only to think of your own optimization of your careers, of how changes in the rating system in the past have altered some of your plans and strategies.

Some systems of measurement clearly have bad features, but tradition, and other niceties, keep them going. For example the state of readiness of a branch of the military. In the Navy ships are inspected on a regular routine, one feature after another, and the skipper gets the ship and crew ready for each one, pretty much neglecting the others until they come up. The skipper scores high, to be sure. But when we face simulating war games, what is the true readiness of the fleet? Surely not what the reports say—as you can easily imagine. But what do we have to use? Of course we must use the reported figures—we would not be believed if we used other data! So we train people in war games to use an idealized fleet and not the real one! It is the same in business games; we train the executives to win in the simulated game, and not in the real world. I leave it to you to think about what you will do when you are in charge and want to know the true readiness of your organization. Will random inspections solve everything? No! But they would improve things a bit.

All organizations have this problem. You are now at the lower levels in your organization and you can see for yourself how things are reported and how the reports differ from reality—it will still be the same unless you, when you are in charge, change things drastically. The Air Force uses what are supposed to be random inspections, but as a retired Navy Captain friend of mine once observed to me, every base commander has a radar and knows what is in the air and if he is surprised by an inspection team then he must be a fool. But he has less time to prepare than for scheduled inspections, so presumably the inspection reports are closer to reality than when inspections only occur at times known far in advance. Yes, inspections are measurements, and you get what you measure. It is often only a little different in other organizations—the news of a coming measurement (inspection) gets out on the grape vine of gossip, and the receivee, while pretending to be surprised, has often prepared that very morning for it.

Another thing which is obvious, but seems necessary to mention; the popularity of a form of measurement has little relationship to its accuracy or relevance to the organization.

Still another thing to mention is all up and down the organization each person is bending things so they themselves will look good—so they think! About the only thing which saves top management is the various lower levels can each only bend things a bit, and often the various levels have different goals and hence the many bending of the truth tend to partially annul each other due to the weak law of large numbers. If the whole organization is working together to fool the top, there is little the top can do about it. When I was on a Board of Directors I was so conscious of this I frequently came either a day early or else stayed a day late, and simply wandered around asking questions, looking, and asking myself if things were as reported. For example, once when inventory was very high, due to the change in the line of computers we were producing which forced us to have parts of both lines on hand at the same time, I walked along, suddenly turned towards the supply crib, and simply walked in. I then eyed things to decide if, in my own mind, there was any great discrepancy or were the reported amounts reasonably accurate.

Again were the computing machines we were supposed to be shipping actually on the loading dock, or were they mythical—as has happened in many a company? Nosing around I found at the end of each quarter the machines to be shipped were really shipped, but often by the process of scavenging the later machines on the production line, and hence the next few weeks were spent in getting the scavenged machines back to proper state. I never could stop that bad habit of the employees, though I was on the Board of Directors! If you will but look around in your organization you will find lots of strange things which really should not happen, but are regarded as customary practice by the personnel.

Another strange thing that happens is what at one level is regarded as one thing, is differently regarded at a higher level. For example, it often happens the *evaluations of capability* of the organization at one level are interpreted as probabilities at a higher level! Why does this happen? Simply because the lower level cannot deliver what the higher one wants and hence delivers what it can do, and the higher level willfully, because it wants its numbers, chooses to alter the meaning of the reports.

I have already discussed the matter of life tests—what can be done and what is needed are not the same at all! At the moment we do not know how to deliver what is needed; reliability for years of operation at a high level of confidence for parts which were first delivered to us yesterday. That problem will not go away, but a lot can be done to design into things the needed reliability. One of my first problems at Bell Telephone Laboratories was the design of a series of concentric rings of copper and ceramic such that for the choice of the radii, as temperatures changed, the ceramic would always be in compression and never in tension where it has little strength. The design has a degree of reliability built into it! Too little has been done in this direction in my opinion, but as I remarked before, when they said there was no time to do it, “There is never time to do the job right, but there is always time to fix things later”.

There are rating systems that have built into them a degree of human judgment—and that sounds good. But let me tell you a story which made a big impression on me. I had produced a computing machine method of evaluating the phase shifts from the measured gains at various frequencies in a signal which replaced a human, hand method. I am not claiming it was better, only the hand method could not do the new job when we passed from voice to TV band widths. A smart man said to me one day, “Before, when humans did things, we could not make further improvements because of the random human variations; now that you have removed the random element we can hope to learn things which were not apparent before”. Methods of rating that do not have human judgment have some advantages—but do not conclude I am against putting in an element of human judgment. Most formal methods are necessarily finite, and the complexity of reality is almost infinite, hence human judgment, wisely applied, is often a good thing—though, as just noted, in a way it stands in the path of further progress with its subjective aspects.

From all of this please do not conclude measurement cannot be done—it can clearly can—but the question of the relevance and effects of a form of measurement should be thought through as best you can *before* you go a head with some new measurement in your organization. The inevitable changes that will come in the future, and the increasing power of computers to automatically monitor things, means many new measuring systems will come into use— ones you yourself may have to design, organize, and install. So let me tell you yet another story of the effect of measurement.

In computing, the programming effort is often measured by the number of lines of code—what easier measure is there? From the coder's point of view there is absolutely no reason to try to clean up a piece of code; quite the contrary, to get a higher rating on the productivity scale there is every reason to leave the excess instructions in there—indeed include a few "bells and whistles" if possible. That measure of software productivity, which is widely used, is one of the reasons why we have such bloated software systems these days. It is a counter incentive to the production the clean, compact, reliable coding we all want. Again, the measure used influences the result in ways which are detrimental to the whole system! It also establishes habits which at a later time are hard to remove.

When your turn comes to install a measuring system, or even comment on one someone else is using, try to think your way through to all the hidden consequences which will happen to the organization. Of course, in principle, measurement is a good thing, but it can often cause more harm than good. I hope the message came through to you loud and clear:

You get what you measure.