Numeracy
Advancing Education in Quantitative Literacy

Volume 3 | Issue 1 Article 10

2010

Parts of the Whole: Observing the State of the System

Dorothy Wallace
Mathematics Department, Dartmouth College, Hanover NH, dorothy.wallace@dartmouth.edu

Recommended Citation
Wallace, Dorothy. "Parts of the Whole: Observing the State of the System." *Numeracy* 3, Iss. 1 (2010): Article 10. DOI: http://dx.doi.org/10.5038/1936-4660.3.1.10

Parts of the Whole: Observing the State of the System

Abstract

This column draws on the approach of statistician J. Edwards Deming to analyze sources and consequences of variation in an education system. Educational systems are not immune from the effects of poor statistical control, which makes it difficult for teachers to teach effectively and for managers such as principals to improve on school performance. It is also argued that the need for statistical control in these areas is in tension, if not outright conflict, with our goals for educating students.

Keywords

quantitative literacy, system, variation, management

Parts Of The Whole

A Column by D. Wallace

The problem of how best to improve the numeracy of a society is a thorny one, embracing the learning process of a single student but rising in scale to include the management and alteration of an entire system of education. With the issue of quantitative literacy always in mind, this column considers various aspects of the systemic workings of education, the forces acting on classrooms, teachers and students, and mechanisms of both stasis and change.

Observing the State of the System

This column owes a huge debt to the ideas and writings of J. Edwards Deming, the statistician whose astute analysis of business and manufacturing helped to put Japan's economy back on its feet after World War II. To business people who may feel his approach is somewhat outdated, one can only say that an approach based on the theorems of mathematical statistics is as permanently useful as those theorems. One benefit of a mathematical approach to a problem is that theorems, once proven, remain true forever. We are still using Newton's calculus and the Pythagorean theorem today, and Deming's observations on manufacturing and business management will remain true for the indefinite future, because they are mathematically based. In this column we will look at some of the basic statistical principles described by Deming and try to see how they might be useful to educators and policymakers.

Deming's idea was quite straightforward. He proposed studying a system in order to identify sources of variation within it, tracing this variation as it propagates through the system. His ideas are particularly useful for those parts of education striving toward the goal of homogeneity, such as meeting literacy and math standards. Notice that this basic approach is quite different from merely giving a standardized test at the end of several years work and drawing conclusions from it. Rather, it is designed to probe into the processes of the system to see how variation is reduced, increased, or otherwise affected as elements (in this case, students), pass through it. Only by understanding the actual processes of the system in this light can one hope to steer it intelligently.

To give an example straight from the standard Deming repertoire, consider the manufacture of screws. The input is some kind of metal, which may vary somewhat in strength and pliability within a single lot. This spread is a cause of variation in the final product, which will be a stronger or weaker screw as a result.

The manufacturing process machines screws, but it has some variation built into it also. Even if the material coming in is identical, there is going to be some (preferably very small) variation in the size of the screw coming out, how deeply the spiral is cut into the shaft of it, and how deeply the slot on top is cut. All of these factors affect the strength and usability of the screw. So, there are at least four identifiable sources of variation contributing to the quality of the final product. An analysis would also take into account human contributions to the process, and how they vary and affect the product.

The first step in analyzing this simple system is to take repeated samples of unfinished screws at each point in the manufacturing process. For example, there will be a point at which a screw has been cast but not yet cut or pressed into final form. One can measure variation in weight, length, and strength of a collection of screws at that point. If one plots the result of average measurements of successive samples over time, one typically sees a picture that looks like Figure 1.

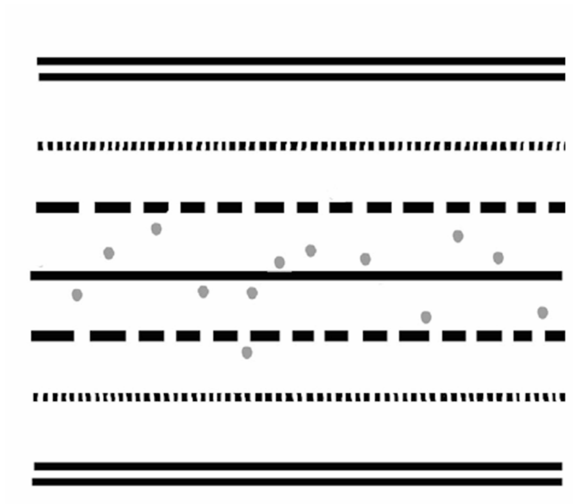

Figure 1. Shewhart control chart of a system in statistical control.

Figure 1 shows a system that exhibits “statistical control”. There is a precise mathematical formulation for this idea, but the translation into English is quite easy. The successive measurements are scattered evenly on either side of the solid line, which is the mean of all of them. A predictable proportion of these measurements (over time) must fall within the dashed lines. A predictable, larger proportion of measurements must fall within the dotted lines. Hardly any measurements should ever fall outside the double lines. A system in statistical control is desirable, because the mean outcome and the variation from that outcome are completely predictable.

DOI: http://dx.doi.org/10.5038/1936-4660.3.1.10

Notice that this chart, called a "control chart," may have to do with standards set externally, but for manufacturing processes it should actually describe what the system is made to do when functioning properly. The range of outcomes delineated by the various lines depends mathematically on the actual variation among screws at this point in the process and also on the size of the sample taken. The distance between any two lines in this chart is sometimes referred to as the "standard error". All of the mathematics involved in constructing this picture is well within the province of introductory statistics.

Education is a process. It is less orderly than a factory, but every small part of it has an input (students who know a little bit less) and an output (students who know a little bit more). We teach students in large groups out of necessity, tracking the mean and variance of the performance of these groups. All of the statistical considerations relevant to factory production are also relevant here. Teachers measure the system regularly using examinations, as do administrators who depend on standardized tests across whole schools, districts and states. At the college level, student evaluations of teachers also produce data that track how the system is working.

Figures 2 and 3 show two examples of control charts where the system is not in statistical control. In the first example, there is too much variation from the mean. In the second example, the mean is slowly creeping upward.

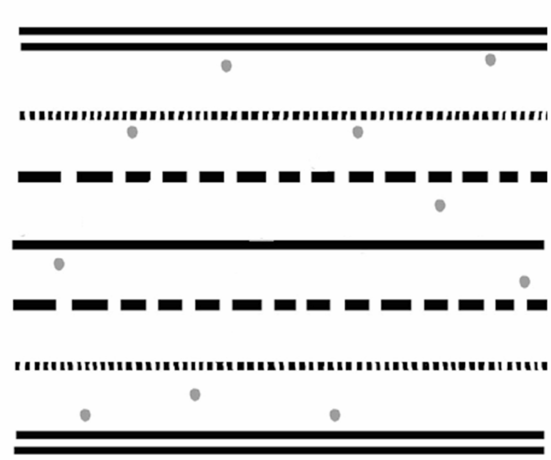

Figure 2. Shewhart control chart of a system that is not in statistical control due to large variation from the mean.

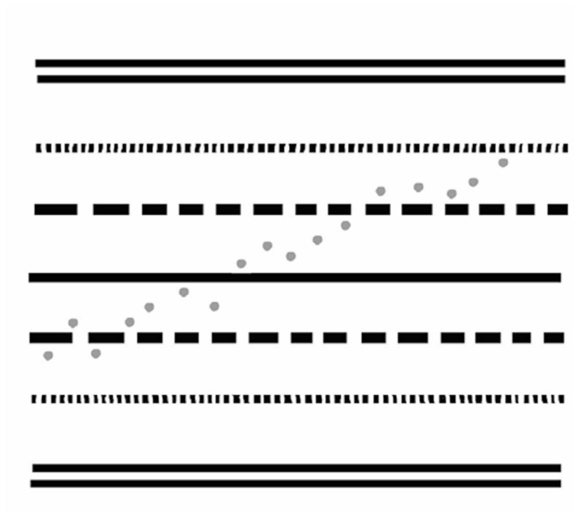

Figure 3. Shewhart control chart of a system that is not in statistical control due to a rising mean.

Figures 1, 2, and 3 illustrate a fundamental tension in the educational process. When an experienced teacher teaches the same material repeatedly, that teacher gets a sense of how to test students efficiently. The teacher learns to make tests that distinguish among students, with a predictable mean performance and standard error from that performance. In order to tell students apart and give a spread of grades, that mean performance is designed to fall short of perfection. As a statistician once remarked to me, any good teacher should be able to design a test where the mean is 75%. On the other hand, a test that results in a distribution like that shown in Figure 2 makes it very easy to distinguish between two groups of students—those falling far above the mean and those far below it. But more importantly, we should ask ourselves what it would mean to be an excellent teacher. If we are honest, we have to admit that what we really desire is for everybody in the class to score perfectly on whatever fair test we can devise. And over the course of years, we would hope that the experienced teacher figures some of this out and produces class scores that resemble Figure 3, in a state of constant improvement. But as we shall see, someone with the responsibility of dealing with the system as a whole will find it easier to manage that system if its inputs and outputs resemble the chart in Figure 1.

A system in statistical control is easier to adjust and improve because of its predictability. Furthermore, the deviation from the mean observed at each stage of the process is a normal part of the system, and is unlikely to derive from any aspect of the system associated to the place where the sample is taken. In other words, if occasionally there is a particularly good screw or a particularly bad one, it is unlikely to be the result of just the last step of the process, or the worker who runs it. Deming calls such variation “common.” When the system is in statistical control, the only way to improve it is “systemically” through study and large-scale change.

On the other hand, if a system is in statistical control up until a particular point in the process, and then proceeds to go out of control at the next step, then the cause most likely lies right nearby, in that step. Deming calls such a cause “special.” It is special precisely because it is not the result of the system as a whole, but some particular local effect. The control charts and the math for making sense of them give management a powerful technique for seeing how to make changes intelligently. If the only measurements taken are at the very end of the process, and if the system is not in statistical control, there is no way to figure out what is causing it. Furthermore, a system that is not in statistical control is much harder to adjust than one that is.

Here is an example from education. Suppose a fourth grade teacher works in the same institution for twenty-five years. Each year, let us say, there is an incoming class of twenty students and at the end of the year they take a battery of exams, receiving scores in math and reading and other things too. Suppose we pick a particular exam score (say, reading) and plot the average score for each year’s class in sequence on a control chart.

The lines on the chart are found using, let us say, data for all test takers in the state. If the chart looks like Figure 1, then the system is in statistical control. Some years, this teacher’s students will fall above the state average. Some years they will fall below. But there is no special cause; all of the variation is of the sort that is natural to any system.

But some years we may say the teacher has done a very good job and others we may criticize. The principal may decide to have a serious formal talk with the teacher in each of the years that the students fall below the state average: about half of all years. The teacher may feel obliged to explain at length that this is exactly what is meant by the word “average”, but to no avail. The teacher may point out in frustration that the incoming classes also display variation and the teacher cannot completely erase it during a single year. The principal, a manager of an educational system, needs to understand that no amount of adjustment on the part of this teacher is going to erase variation that is common to the system.

On the other hand, if the data are found to fail to be in statistical control, the principal should look closely at the situation. Was there anything special that happened in the particularly good or bad years? Perhaps the incoming students should be tested over a period of years to see if the pattern has been inherited from a previous grade. If so, then the fourth grade teacher cannot reasonably be expected to reverse it.

Deming concocted a charming experiment to illustrate the point about statistical control and the “common cause.” He calls it the “red bead experiment.” Recruiting members of his audience to work for him, Deming instructs his team of workers to draw samples of beads, using a fairly random method, from a container of mixed red and white beads. No more than a certain number of red beads are allowed to be in the sample. Of course, the proportion of red beads in a sample is going to be, roughly, the proportion of red beads in the container, give or take some normal statistical variation. After a few trials, it is duly noted that some workers are performing below average, drawing somewhat more red beads than the allowed number. If the allowed number really is in the same proportion to the sample as in the original container, then about half the workers will be found wanting. Deming liked to have a slightly high proportion of red beads in the container, so that failure is almost guaranteed. These workers are eventually fired and new workers hired (from the audience). The experiment is repeated, more workers fired and hired, and the results (in terms of average number of red beads drawn and statistical variation from this mean) never change. The management has treated a common cause as if it were special, acted accordingly, and improved nothing. The "workers", volunteers from the audience, can get somewhat upset about the experience, even though it is only pretend. The whole effect is heightened by Deming playing the part of the sincere but stern manager, giving pep talks to his workers, putting motivational posters and slogans up on the walls, and chastising those who perform poorly.

Policy decisions are systemic in nature; they attempt to readjust the system in order to improve the output of it in some way. But a system that is not in statistical control cannot be aimed. To illustrate this point, consider Figures 4 and 5.

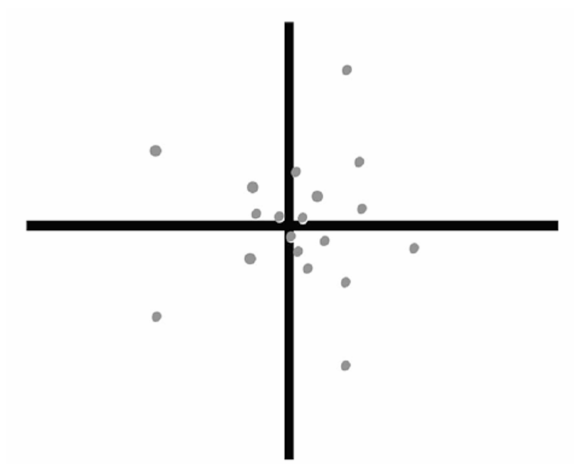

Figure 4. Two-dimensional data exhibiting statistical control, a good shotgun.

DOI: http://dx.doi.org/10.5038/1936-4660.3.1.10

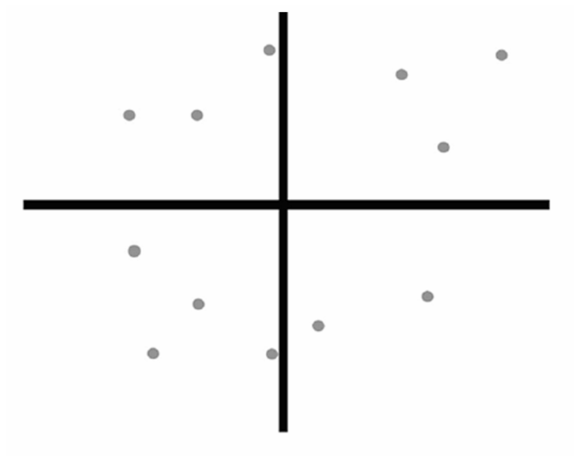

Figure 5. Two-dimensional data with poor statistical properties, a bad shotgun.

Figure 4 is the kind of image produced when the sites are set on a gun. The gun is fired repeatedly at a target until a statistical picture emerges of what it does. The sites are set to aim at the intersection of the axes, and the data points form a distribution centered at that point. The problem of aiming the gun is clear now. What is really being aimed is the center of this distribution. Furthermore, the gun is a system in statistical control. If you aim it at something, a large proportion of shots will land close to that something.

Figure 5 is a hypothetical output of another gun. This gun is not in statistical control. Although there is an average location where shots land, the shots do not usually fall near the average. In fact, the shots consistently fall very far from the average location. It is impossible to aim this gun. No matter where one points it, a large proportion of shot is fated to fall very far away from the target. This gun is behaving like a two-dimensional version of the control chart in Figure 2.

When educators and policy makers speak of the widening gap in student performance, they are describing an educational system that is behaving rather like the gun in Figure 5. Unless the system can be brought into statistical control, the effect of any intervention is likely to be capricious. Deming is at pains to point out in his work that it is nearly impossible to steer the mean outcome of a process until that process is brought into statistical control. Successful efforts to control variation typically precede successful efforts to improve overall performance.

Schools are not factories and students are not products. Yet, when a sizeable high variance group of students is attempting to learn a high priority subject, the same problems emerge. A high priority subject, as I am using the term now, requires a lot of intervention on the part of the teacher (see column 2 in this series for an explanation of this statement). If the class is large, the teacher is forced to interact with it as a unit, as there is not time to teach subpopulations separately. If the background and preparation of these students comes from a system in statistical control, the teacher can aim lessons at the mean of the population, knowing that this will benefit the most students. The teacher also knows that some will have to do extra work to stay caught up and that others will find the material too easy. An experienced and dedicated teacher may develop strategies for handling the extremes in the distribution.

However, if the background and preparation of the students looks like the data in Figure 5, then the teacher has nowhere to aim the lesson. To create a lesson for a hypothetically average student is to create a lesson for nobody, because there is nobody in the class that comes close to that average. It is hard to design a lesson intentionally that would be equally incomprehensible to everyone in the class, yet that is exactly what the mean value of the data in Figure 5 represents. To create a lesson that is perfect for one clump of students is to serve many others very poorly, even worse than aiming at the mean. Treating students as unique individuals to be taught separately based on their particular background (which is not usually feasible in any case) is a strategy that intentionally removes any pretence of a system in control.

Ironically, in a an orderly educational system in statistical control, society and educators will label students performing at the mean level as being not particularly talented in that subject, even though the system is designed to place their performance near the mean. Here is the main point of contrast with the example of a factory. A screw whose measurements fall exactly at the mean value of the control chart is a perfect screw. A student whose measurements fall exactly at the mean value of test measurements is a mediocre student. This observation is at the root of all difficulties with educational standards and assessments of those standards, and it highlights the judgmental and emotional background against which all management decisions based on assessment instruments must be made.

DOI: http://dx.doi.org/10.5038/1936-4660.3.1.10