MONITORING TEACHERS’ PERFORMANCE USING STATISTICAL PROCESS CONTROL CHARTS

by

Ben A. Maguad

Andrew University, Berrien Springs, Michigan, U.S.A.

ABSTRACT

The use of statistical process control charts to monitor teachers’ performance will help the academic institution to understand which problems are attributable to the system and which ones are attributable to special causes. It will help the administration to discover which teachers are performing outside the limits of variation of the system on either the good side or the poor side and which ones are performing within the calculated limits of differences attributable to the system so that appropriate responses can be crafted to address the needs of different groups of teachers. This paper presents how a statistical process control chart can be developed from teachers’ performance evaluation data, and how this chart can be used to monitor a teacher’s performance over a period of time.

KEY WORDS: process variation, common cause variation, special cause variation, statistical process control, control chart for monitoring performance of teachers.

INTRODUCTION

During the last few decades, evaluation of teaching has changed significantly in terms of purpose and methodology (Ory, 2000, p. 13). Traditionally, evaluation information was used primarily to meet student demands for public accountability. Later, evaluation was done to help faculty improve and develop performance. In recent years, however, teaching evaluation seemed to have been driven by the administrative desire to meet the demands of the public for institutional and legal accountability. Teaching evaluation has evolved from a simple reliance on the supervisor’s assessment to a more formal, systematic, and multiple approaches, which include student ratings, peer reviews, self-evaluations, document reviews, and evidence of achievement. This set of methods was developed to meet the needs for sophisticating the evaluation tools for monitoring teachers’ performance and for providing administrators tangible and objective information.

Performance evaluation of teaching is normally employed to assess the performance of teachers and generate information about his or her productivity at work. The conventional evaluation process normally involves the department chair or the supervisor (acting unitarily or jointly with his or her subordinate) setting objectives for a certain period of time. These objectives might focus on developing knowledge, skills, results (such as output and productivity), or behavior. The evaluation activity often involves the use of a performance-rating questionnaire followed by a supervisory review of accomplishments, strengths and weaknesses, and other characteristics related to the job being appraised. This process may be followed by announcements of raises, bonuses, or promotions.
It is expected that as new purposes and audiences are added, the methods to evaluate teaching would also change. But will these changes help to improve the central purpose of the evaluation, which is to help teachers continually improve and develop so that the student learning experience in the classroom and in the school can be enhanced?

CONVENTIONAL PERFORMANCE EVALUATION: SOME CONCERNS

Conventional performance evaluations tend to “encourage short-term performance at the expense of long-term planning, discourage risk taking, build fear, undermine teamwork, and pit people against each other for the same rewards” (Walton, 1986, p. 91). As a result, people work for themselves at the expense of the institution – behavior that encourages destructive competition, in-fighting and turf building. Deming (1986, p. 102) equates conventional performance evaluations with “management by fear” and asserts that such leave “people bitter, crushed, bruised, battered, desolate, despondent, de-jected, feeling inferior, some even depressed, all unfit for work for weeks after receipt of rating, unable to comprehend why they are inferior. It is unfair, as it ascribes to the people in a group differences that may be caused totally by the system that they work in.” Conventional performance evaluations create a shortage of winners within the educational institution and do little to motivate teachers or enhance their satisfaction.

The following objections, as applied to education, have been made regarding the use of conventional evaluation systems (Milliman and McFadden, 1997, p. 67; cited in Evans and Lindsay, 1999, pp. 315-316).

- They tend to foster mediocrity and discourage risk taking.
- They focus on short-term and measurable results, thereby discouraging long-term planning or thinking and ignoring important behaviors that are more difficult to measure.
- They focus on the individual and therefore tend to discourage or destroy teamwork within and between departments.
- The process is detection-oriented rather than prevention-oriented.
- They are often unfair, since evaluators frequently do not possess observational accuracy, and
- They fail to distinguish between factors that are within the teacher’s control and system-determined factors that are beyond his or her control.

UNDERSTANDING PROCESS VARIATION

All work occurs in a system of interconnected processes, which contains many sources of variation. The complex interaction of materials, tools, machines, operators, methods, and the environment generate variations in outputs of the production process. The combined effect is stable and can usually be predicted using statistical methods like the control chart. This type of variation appears at random and can be attributed to chance or to common causes. A process that is being affected only by chance causes of variation is said to be stable and in a state of statistical control (Lewis and Smith, 1994, p. 46). The process output is therefore predictable within a range or band. Common causes are the “myriad of ever-present factors (for example, process inputs or conditions) that contribute in varying degrees to relatively small, apparently random shifts in outcomes day after day, week after week, month after month” (Roehm and Castellano,
They generally account for about 80 to 95 percent of the observed variation in the
output of a production process (Evans and Lindsay, 2005, pp. 518-519). Common cause variation
can only be reduced if the product (or service) is redesigned, better technology employed, or bet-
ter training provided for those who participate in the production process. It represents the great-
est opportunity for long-term improvement.

The other type of variation that may occur in a production process is the result of special
causes, also called assignable causes of variation. Special causes can be attributed to external
sources that are not inherent in the process. They are the product of special circumstances, a
temporary glitch in the system (Gabor, 2000, p. 189). Some examples of special causes are a bad
batch of material from a supplier, a poorly trained substitute machine operator, a broken or worn
tool, or faulty calibration of measuring instruments. Since special causes are unpredictable and
can have a relatively major effect on the system, they must be identified and removed. When a
process is affected by special causes, it is out of statistical control. Special causes can be identi-
fied and eliminated by workers who have been properly trained to analyze the process.

Variation exists in every operation and in every process including human performance.
Applied to our total educational system, it follows that teachers cannot be expected to perform at
the same level. There will always be differences. The question is, what do the differences mean?
Whether they mean something or not mean anything at all depends on the outcome of the control
chart analysis. Other features of performance evaluation that stem from failure to understand
variation from common causes are ranking and merit pay. Ranking creates competition between
people, schools, divisions, and teams. It demoralizes employees (teachers) and produces a Pyg-
malion effect on them. When one is rated high at the start, he or she tends to stay high. The con-
verse is also true. When one is rated low at the start, he or she tends to stay low (Deming, 1993,
p. 27). When people are placed in categories and labeled “above average,” “average,” and “be-
low average,” the labels tend to stick (Huge, 1990, p. 82). The results of performance evaluations
are always the same: half of those rated will be above average and half will be below. How do
we judge the performance of an educational institution where half of the workers are below aver-
age?

The following equation will help to explain the futility of attempts to rank teachers in our
schools. Let \( a \) be the contribution of the individual teacher and \( ba \) the (random) effect of the
system on his or her performance. Suppose we assign some number for his or her apparent per-
formance, such as ten mistakes during the year or a revenue contribution of $10,000. The above
relationship can be represented by the following equation:

\[
a + (ba) = 10
\]

From an algebraic standpoint, it will not be possible to solve for \( a \), since there are two
unknown variables, but only one equation. Unfortunately, those who rank teachers or use the
merit system think that they can solve for \( a \). The other predominant term \((ba)\), which incorpo-
rates the effect of the system on the worker’s performance, is often ignored. This is precisely the
reason Deming (1986) was strongly opposed to traditional performance appraisal systems be-
cause they fail to distinguish between factors that are within the teachers’ control and system-
determined factors that are beyond their control (cited in Evans and Lindsay, 1999, p. 316). Most
of the performance problems in any organization may be attributable to the system, not the people.

THE PROCESS CONTROL CHART

A control chart is a run chart to which two horizontal lines, called control limits are added (see Figure 1 for the original structure of a control chart in manufacturing). The two lines are the upper control limit (UCL) and the lower control limit (LCL). Control limits are chosen statistically to provide a high probability (generally greater than 0.99) that points will fall between these limits if the process is in control. They make it easier to interpret patterns in a run chart and draw conclusions about its state of control.

If sample values fall outside the control limits or if nonrandom patterns occur in the control chart, then special causes may be affecting the process. Thus, the process is not stable. The process should be examined and corrective action taken as appropriate. In manufacturing, the process mean is ordinarily derived by averaging the performance of machines used in production. Over time, this average produces a standard against which the UCL and LCL are calculated to determine whether performance data fall within the control limits (derived from common causes) or outside of them (derived from special causes). Thus if a machine shows considerable performance above or below the UCL and LCL, that machine needs to be fixed or replaced because it is creating defective products. Manufacturing applications of control charts rely exclusively on quantitative data.

<table>
<thead>
<tr>
<th>Performance</th>
<th>Measurement Over Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCL</td>
<td>Upper control limit</td>
</tr>
<tr>
<td>Process</td>
<td>Center line</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>LCL</td>
<td>Lower control limit</td>
</tr>
</tbody>
</table>

![Figure 1. Structure of a Control Chart for Manufacturing](image)

The chart in Figure 1 can be modified for applications in education (see Figure 2). The center line or the process mean may be interpreted as the average teachers’ performance (ATP) based on a designated criterion (e.g. teaching, publication, service, etc.). Instead of UCL and LCL we use the upper limit of teachers’ performance (ULTP) and the lower limit of teachers’
Performance (LLTP). ULTP and LLTP are not really for “control” since performance excellence needs to be identified, understood, and perhaps used to raise the process mean to achieve continual improvement. Performance below the LLTP needs to be identified so those teachers get additional training and education.

<table>
<thead>
<tr>
<th>Performance</th>
<th>Measurement Over Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULTP</td>
<td>Upper limit of teacher’s performance</td>
</tr>
<tr>
<td>ATP</td>
<td>Average teacher’s performance</td>
</tr>
<tr>
<td>LLTP</td>
<td>Lower limit of teacher’s performance</td>
</tr>
</tbody>
</table>

**Figure 2. Structure of a Control Chart for Evaluating Teachers Performance**

Performance ratings of teachers in schools using a numerical system based either on a single measure or on a weighted index tend to divide them into three groups:

1. Teachers who are outside the “control” limits (above the ULTP) on the good side.
2. Teachers who are outside the “control” limits (below the LLTP) on the poor side.
3. Teachers who are between the “control” limits (between the ULTP and the LLTP).

There may not be anybody in Group 1 or 2. However, if there is any, then that teacher requires individual attention. We usually have a pretty idea of who the superstars are and also who the misplaced ones are. Knowledge of control charts (statistical process control) teaches that those in Group 3 (people between the “control” limits), which comprise the majority of the teachers, must not be ranked. The primary reason for this is that differences between levels (or points) within the “control” limits come from the system itself, not from the teachers. There is no justifiable distinction between them. The differences come from the system and therefore must be ascribed to the system and not to the teachers themselves.

One common practice that can be benefit from the use of statistical process control is the student evaluation of teaching, which is administered regularly in schools during a given term (quarter, semester, or a year). Students are asked to evaluate their teachers and their classes on Likert scale (Likert scale (1 = poor; 5 = excellent) using a standardized survey containing a set of itemized questions plus perhaps one or two global questions that focus on the teacher and/or class as a whole. Statistical reports are usually prepared by computing the class average for each question and also the group average (the group may be a department, an entire school/college, or the entire institution). It is common, when analyzing the statistical figures, to compare an individual performance against the group norm. But what does it really mean when one’s performance is below or above the average? When an average is computed, about half of the group is
expected to be above the average and half below. How do we account for the fact that half of our workers are below average? How would one feel if he or she is branded a below-average performer? Instead of relying too much on averages it may be better to draw a control chart (e.g. mean chart: $\nu$-chart; range chart: $R$-chart) for all members of the group and establish upper and lower “control” limits. This can be done for several quarters, semesters, or years to establish a clear pattern and determine who are consistently outside the control limits (if there is any) and who are within them. Those whose evaluation ratings are within the “control” limits must not be ranked since differences among them can only be attributed to the system. Those who are consistently outside the “control” limits on the bad side (below the LLTP) may need special help. Those who are consistently outside the control limits on the good side (above the ULTP) may deserve some special recognition. But this should be approached with caution because the life of a student in an educational institution is affected by many individuals. This means that no single person can be held responsible for whatever a student does (or fails to do) in the future. As a matter of fact, the ultimate outcome of a student’s education is not known until many years later. Therefore, to limit the reward to some teachers, when their performance is part of a larger system and their effectiveness is not known until many years later, may be unjustifiable (Rinehart, 1993, p. 90).

The following is a numerical example of a control chart that is drawn based on ratings received by ten faculty members from a school during a given semester. The ratings were based on a Likert scale of 1 to 5 (1 = poor; 2 = fair; 3 = good; 4 = very good; 5 = excellent). They were obtained in response to the following question:

*Independent of the course, what is my overall rating of this instructor’s effectiveness?*

<table>
<thead>
<tr>
<th>FACULTY</th>
<th>CLASS SIZE (n)</th>
<th>DATA ($\bar{x}$ = mean)</th>
<th>ULTP</th>
<th>ATP</th>
<th>LLTP</th>
<th>DATA (R = range)</th>
<th>ULTPR</th>
<th>ATPR</th>
<th>LLTPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-1</td>
<td>15</td>
<td>4.00</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-2</td>
<td>12</td>
<td>3.91</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-3</td>
<td>14</td>
<td>4.33</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>1</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-4</td>
<td>11</td>
<td>3.91</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-5</td>
<td>15</td>
<td>4.00</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-6</td>
<td>15</td>
<td>4.00</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-7</td>
<td>12</td>
<td>3.91</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-8</td>
<td>14</td>
<td>4.33</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-9</td>
<td>11</td>
<td>3.91</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
<tr>
<td>F-10</td>
<td>15</td>
<td>3.80</td>
<td>4.49</td>
<td>3.99</td>
<td>3.49</td>
<td>2</td>
<td>3.39</td>
<td>2.00</td>
<td>0.61</td>
</tr>
</tbody>
</table>

$$I = 13 \quad A_2 = 0.249 \quad D_4 = 1.693 \quad D_3 = 0.307$$

Table 1. Student evaluation of teaching statistics
INSTITUTE LEADERSHIP

Most quality experts agree that quality is determined in the executive suite. The quality of the output of the organization cannot be better than the quality determined at the top. Deming (1993, p. 35) states that in his own experience, most troubles and most possibilities for improvement (about 94%) are attributable to the system. Only about 6% are attributable to special causes. Improving the educational system is the responsibility of administration. No amount of care by teachers or their skill in workmanship can overcome the fundamental faults of the system. It is the responsibility of leadership to understand and improve the system.

A good leader will strive to discover who if any of his/her teachers performs (a) outside the system on the good side, (b) outside on the bad or the poor side, and (c) within the calculated limits of differences to be ascribed to the system. Anyone whose performance is outside the limit of variation of the system on the good side may deserve some recognition. There is rational basis to predict that he or she will perform well in the future. He or she could be a focal point for teaching other faculty members how to do their job better. If a faculty member continues to hold a superior relative position through at least seven successive time periods (Deming, 1986, p. 116), we may safely conclude that he or she is indeed superior based on the index of performance used.

Faculty members, who are on the poor side of the system, if there is any, will require individual help. A leader should spend time with everyone in this group to find out what kind of help they need. For some, it may be additional training. For others, it may be just be counseling to help restore confidence and performance. And for some others, it may just be an offer to move to a more suitable position within the organization or a recommendation to find a more suitable job elsewhere. Or perhaps it may just be a simple trip to the doctor (e.g. physician, oculist, and dentist) for health checkup and treatment.
A leader should not rank faculty members whose performance falls within the calculated limits of differences attributable to the system (within the “control limits”). Apparent differences between members of this group arise almost entirely from the action of the system, not from the faculty members themselves. These differences may not mean anything at all. There is no evidence that the performer with the highest rating in this group will in the future be a better performer than the one who had the lowest rating. The leader has the responsibility to improve the system, to accomplish ever greater and greater consistency of performance so that these apparent differences continually diminish.

A common fallacy is the supposition that it is possible to rate people; to put them in rank order of performance for next year, based on performance last year. The performance of anybody is the result of a combination of many forces – the person himself, the people that he works with, the job, the materials that he works on, his equipment, his customer, his management, his supervision, environmental conditions (noise, confusion, poor food in the company’s cafeteria). These forces will produce unbelievably large differences between people. In fact ..., apparent differences between people arise almost entirely from action of the system that they work in, not from people themselves. A man not promoted is unable to understand why his performance is lower than someone else’s. No wonder; his rating was the result of a lottery. Unfortunately, he takes his ratings seriously (Deming, 1986, p. 109-110).

CONCLUSION

Many questionable practices in our institutions arise from failure to understand the difference between common causes and special causes of variation. A proper understanding of process variation will help us to understand which problems are attributable to the system and which are attributable to special causes. People should not be ranked. Judging people, putting them into slots, does not help them do a better job. Forcing competition between them when in reality the dominant cause of variation is due to the system will just lead to conflict and not improvement. Instead, the administration should manage the institution as a system so that the function of every component (schools, colleges, departments, etc.) contributes to the optimization of the system. It should encourage communication and cooperation among components and give each one a chance to take pride in his or her work. The administration should determine who is in need of special help and make sure he or she receives it. It should not differentiate between the above average and the below average. Instead, it should actively work to determine the real causes of problems and then work to eliminate them. A good administrator forgives mistakes and fosters cooperation among institutional components. Most of all, a good administrator fosters the continual development of all those who work for him or her. If administrators were to spend as much time and effort to improve processes as they currently do on ranking, rating, and rewarding people (individuals, teams, departments, divisions, schools, colleges, etc.) at the top, and punishing those at the bottom, the results would be a tremendous improvement in all aspects of the school operations.
REFERENCES


ABOUT THE AUTHOR

Dr. Ben A. Maguad is an Associate Professor of Management at Andrews University, Berrien Springs, Michigan, U.S.A. He earned his PhD in Business & Management from the University of South Australia, Adelaide, South Australia with focus on the application of quality management in higher education. He worked and taught in colleges and universities in a number of countries including Philippines, Thailand, Hong Kong Special Administrative Region, Trinidad & Tobago, and the United States.