An Assessment of Statistical Process Control-Based Approaches for Charting Student Evaluation Scores

Xin Ding†

PhD Candidate, David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, e-mail: pmgtxd@business.utah.edu

Don Wardell

Associate Professor and Chair, Department of Management, David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, e-mail: mgtdgw@business.utah.edu

Rohit Verma

Associate Professor, School of Hotel Administration, Cornell University Ithaca, NY 14853, e-mail: Rohit.verma@business.utah.edu

We compare three control charts for monitoring data from student evaluations of teaching (SET) with the goal of improving student satisfaction with teaching performance. The two charts that we propose are a modified p chart and a z-score chart. We show that these charts overcome some of the shortcomings of the more traditional charts for analyzing SET data. A comparison of three charts (an individuals chart, the modified p chart, and the z-score chart) reveals that the modified p chart is the best approach for analyzing SET data because it utilizes distributions that are appropriate for categorical data, and its interpretation is more straightforward. We conclude that administrators and faculty alike can benefit by using the modified p chart to monitor and improve teaching performance as measured by student evaluations.

Introduction

Previous research has dealt with the analysis of teaching evaluation, but such analysis has concentrated on exploring the variables that have significant effects on student evaluation of teaching (SET) scores, such as age (Centra, 1993), gender (Feldman, 1977), and education level (McKeachie, 1979). Other research in the education and economic literatures has examined the linkages between evaluation of teaching performance with merit pay, tenure, and promotion (Katz, 1973; Marsh &Dillon, 1980; Siegfried&White, 1973). Most of those studies, however, ignored or discounted a major purpose for analyzing the evaluation data set, which is to effectively monitor and improve instructor performance and student satisfaction levels. A recent exception is Marks and Connell (2003), who proposed the use of an individuals chart to monitor instructor performance. Another exception is Ozgur and Strasser (2004), who

recommended the *t* distribution as the best solution to construct confidence intervals and test hypotheses when the population standard deviation is unknown. In this article, we argue that these approaches are not appropriate for SET data. In searching for the best approach to analyze SET data, we applied the modified p chart (Wardell & Candia, 1996) and a *z*-score chart to course evaluation data for 3 years from the business school in a large university in the United States. Based on the comparison, we recommend the use of the modified *p* chart because it is more appropriate for SET data and because of its ability to identify special causes of variation in instructor performance.

The purpose of our article, therefore, is not to argue the inherent validity (or lack thereof) of student evaluation data. Instead, because such data are ubiquitous and used frequently in the evaluation of teaching effectiveness, we focus on how SET data can be used to evaluate and ultimately improve student satisfaction with instructor performance. In particular, by carefully studying special causes signaled by control charts, instructors and administrators will be able to find ways to meet the educational needs of the students. They may also be able to see when course level factors such as whether the class is required, its level, enrollment size, and so on have an effect on the student evaluation scores.

In the next section, we review the literature on SET data that is relevant to our work. In subsequent sections, we describe and then compare the charts noted above. We end with concluding comments and provide directions for additional work.

Previous Studies

Cashin (1995) had reported that more than 1,500 books and articles dealt with student ratings of instruction. The number increased to 2,000 in 1998 as reported by Wilson (1998). Generally, previous studies found that ratings on the different dimensions of teaching that SET forms measure are positively correlated with student achievement on multiple-choice tests (Herbert & Lawrence, 1997; Sylvia & Phillip, 1997). The validity of SET and the relationships between it and faculty research productivity have also been studied (Kanagaretnam, Mathieu, & Thevaranjan, 2003; Stratton, Steven, & Randall, 1994; Wallace & Wallace, 1998).

Rather than arguing the inherent validity of student evaluation data (Davis, 1995), there is another stream of research that focuses on the analysis of such data. As suggested by Mesak and Jauch (1991), the two leading purposes for faculty evaluation in institutions of higher education today are to improve performance (Brightman, 1987; Evans, 1986) and to provide the rationale for faculty personnel decisions such as merit pay, tenure, and promotion. Mesak (1991) examined how institutions of higher education might operationalize performance evaluation as related to research, teaching, and service and proposed a model that allows coupling performance evaluation and relevant market considerations with merit pay, tenure, and promotional decisions. Brightman, Elliott, and Bhada (1993) adopted an innovative norm report, which provides comparative percentile data on the six factors underlying their 33-item questionnaire. Their study suggested that faculty receiving low percentile scores should review

the questionnaire items which load highly on the factors to develop a self-improvement plan, or ask what others have done to achieve high percentile scores on each factor.

Starting from the quality control perspective, Martin (1998) studied the problems associated with student evaluations and suggested those problems were created by a whole set of flawed assumptions related to the design of the higher education system. Those flawed assumptions include the view that concentrating on the performance of individuals is the key to optimizing the performance of the system, and individual performances are independent and can be measured separately. Based on Deming's theory, he argued that a solution will require systematic change and that faculty and administrators should stop treating each course separately and form cross-functional teams to ensure continuous improvement in higher education systems.

As used in quality management, one way to facilitate process improvement is the implementation of control charts. Such charts provide a way to monitor performance over time and to look for process changes, which are indicated by points outside of the control limits (Evans & Lindsay, 1999). In some processes, points outside the limits indicate that the process has deteriorated, while in other instances such points indicate a possible improvement. By investigating the special causes of points outside of the control limits, process owners can determine how to improve the process in the future.

Using such an approach, Marks and Connell (2003) applied statistical control charts to analyze SET data. They plotted individual evaluation scores and residuals from a regression model, in which the teachers' summary rating is expressed as a function of the student's expected grade. They calculated the control limits for the individuals chart as $\bar{\bar{x}} \pm \frac{2}{3} E \bar{R}$, where $\bar{\bar{x}}$ is the grand mean of all observations examined, E is a control chart factor that depends on the subgroup size, and $\bar{R}$ is the mean of the ranges of subgroups into which the data are divided. The multiplier $\frac{2}{3}$ is used to create tighter limits so as to detect changes more quickly (at the expense of more frequent false alarms). Marks and Connell (2003) concluded that the instructor's performance is unusual only when a rating (or residual from their regression model) is outside of the control limits.

We believe that there are three potential problems with Marks and Connell's (2003) approach. First, their implementation did not include any time element. As noted by Evans and Lindsay (1999), control charts are generally constructed on the basis of a time series of continuous observations. In contrast, the Marks and Connell (2003) approach was essentially a series of hypothesis tests shown graphically over instructors instead of over time. Specifically, Marks and Connell's random sampling assumed that there was no natural time order for the average instructor ratings and they randomly selected samples of size 2 by pooling the data set. In their study, they looked at only between-group variation and completely ignored within-group variation. Although their result shows differences between instructors, it conveys no information regarding how well each instructor performs over time.

Second, use of individuals control charts assumes several things about the data that are not met when SET data are used. For example, Wardell and Candia (1996) described the deficiencies of $\bar{x}$ charts (which are similar to the individuals chart proposed by Marks and Connell (2003)) when applied to customer satisfaction survey data, which SET data essentially are. The main problems are that the data are categorical and hence not normal (which means that the control chart constant multiplying the average range is especially not appropriate for estimating the standard deviation), and the sample sizes are often large (larger than those that are used in control chart constant tabulations) and vary, thereby making implementation of traditional Shewhart charts complicated.

Finally, the Marks and Connell approach appears to ignore class size. Their approach uses mean scores for each course, but then treats the means as individual observations. We believe that the size of the class is important when considering evaluation scores, and hence prefer to use them as part of the input to the control chart analysis. Therefore, in our opinion, the Marks and Connell (2003) approach needs refinement, which we provide in the next section.

Proposed Alternative Control Charts

To overcome the shortcomings of traditional control charts when applied to SET data, we propose two alternatives. These charts can be used on categorical data and are not difficult to implement even when sample sizes are large and vary.

Modified p Chart

Wardell and Candia (1996) pointed out in their research on hospital satisfaction data that nonnormally distributed observations from surveys with categorical scales and unequal sample sizes made it difficult to apply Shewhart charts to survey data. Without forcing normality on the original data, they proposed the modified *p* chart (an extension of the *p* chart), which does not assume the normal distribution (and, in fact, utilizes distributions that are appropriate for categorical data) and allows the control limits to vary with sample size.

For a given question on an SET form or questionnaire, assume that there are m possible responses presented to the student (typically $m = 5$ or 7). Assume initially that the historical or assumed proportion of students who have given response i is known and is p_i, with probability that the response of a given student to the particular question is i, $i = 1, 2, \ldots, m$. Then the response of a given student to the particular question is a discrete random variable with mean (or expected value) and variance of $\mu = \sum_{i=1}^{m} i p_i$ and $\sigma^2 = \sum_{i=1}^{m} i^2 p_i - \mu^2$, respectively. Wardell and Candia (1996) computed the control limits for the modified *p* chart as

$$UCL_x | LCL_x = \sum_{i=1}^{m} i p_i \pm \frac{3}{\sqrt{n_x}} \sqrt{\sum_{i=1}^{m} i^2 p_i - \left(\sum_{i=1}^{m} i p_i\right)^2}$$

where n_i is the number of respondents on the ith SET survey. Because n_i can vary with i the control limits for the modified p chart may also vary. As suggested by Wardell and Candia (1996), if varying limits are undesirable, simplified average limits for the total sample can be calculated based on an average sample size. Such a procedure is sometimes recommended for np chart implementation when sample sizes vary (Gitlow, Gitlow, Oppenheim, & Oppenheim, 1995).

When historical or known values of the p_i values are not available, the estimated p values can be calculated from the data as $p_i = x_i / \sum_{i=1}^{k} n_i$, where x_i is the actual number of students giving response i on SET survey

i

z-Score Chart

Although SET data are not normally distributed, if class sizes are large, we can argue that the sample means may be close to normal. Moreover, it is common for administrators to use mean scores from classes. Marks and Connell used sample means in their procedure. We propose an alternative method for using the mean values, namely the monitoring of z scores.

Any sample mean $\bar{x}$ derived from a population with a known mean and

standard

deviation is easily transformed into a standard normal distribution by computing $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, where $\bar{x}$ is the average rating for the course, μ is the departmental mean for a particular question, σ is the departmental standard deviation for a particular question, and n is the number of students in a particular course. The departmental standard deviation of a particular question is computed as $\sigma = \sqrt{\frac{1}{N}\sum(\bar{x}_i - \bar{x})^2}$, where N refers to the number of courses and $\bar{x}_i$ refers to the mean evaluation for the ith course ($i = 1, \ldots, k$).We can therefore compute z scores for each question on the SET survey and plot them on a control chart. As with the modified p chart, we adopt *3-sigma limits* in this article because three standard errors encompass nearly all the normal distribution. As suggested by Sincich (1986), Younger (1979), and Ronald (1988), an observation is out of control when the absolute value of the z score,

$|z|$,
exceeds 3.

In the following section, we compare the individuals, modified np and p score control charts in monitoring teaching performance.

Application of the Charts

Data analyzed in this article came from student teaching evaluations administered for 3 years within the management department of a large university located in the United States. In this article, we only focus on the question regarding “Overall-Teaching Effectiveness,” which is critical for evaluating instructors’ teaching performance. Once we identify any out-of-control points for the teaching effectiveness, we will generally go back to other questions and seek reasonable explanation for the phenomena. Student responses to the question of “Overall-

Teaching Effectiveness" are based on the following five-point scale: Strongly agree = 5, Somewhat agree = 4, Neutral = 3, Somewhat disagree = 2, and Strongly disagree = 1. The number of students giving each response, the instructor averages and course level, class size, and student gender were recorded for each course. To avoid the problem that arises when composing the Shewhart control charts with insufficient observations, we filtered out those individual instructors who taught fewer than 10 courses across the observation time period.

After filtering out the courses enrolled with fewer than 10 students and those instructors who taught fewer than 10 courses, 158 distinct sections from spring semester of 1995 to spring semester of 1998 were identified for 18 faculty members, with the number of enrolled students varying from 10 to 75. Analysis of the results was straightforward. After

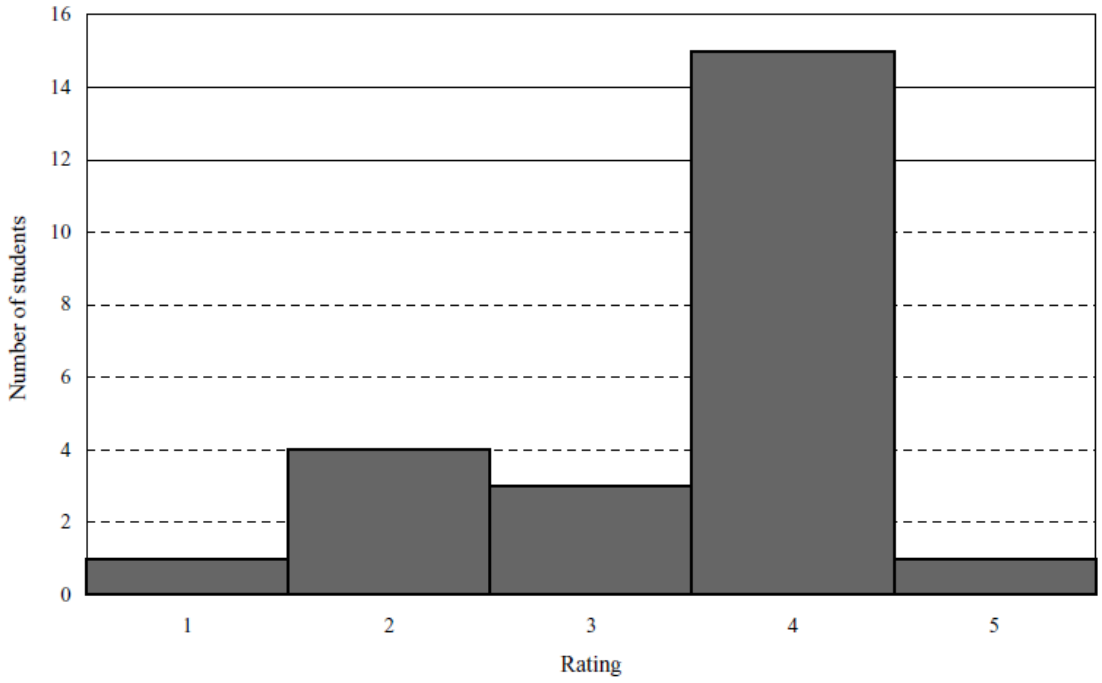

Figure 1: Nonnormal distribution of SET data set.

identifying the 158 distinct courses, each instructor was analyzed at the individual level and the appropriate statistics were computed and charted on the different charts.

Before comparing the three control charts, we present an example of common SET data. Taking the teaching evaluation for the first course by instructor 1 in Fall semester of 1995 as an example, we see that the ratings are not distributed normally with most students rating the instructor at 4 (see Figure 1).With the limited sample size in these data, it is not appropriate to assume that the normal distribution applies (notwithstanding the central limit theorem). Hence, the control chart constants often used in traditional control charts (such as) may not be appropriate for these data.

Individuals Chart

Following Marks and Connell's approach, we constructed the individuals charts for single evaluation means by pooling the data set and randomly selecting samples of size 2. In the present application, the 158 average instructor ratings were pooled, and we used Gauss to program and generate 1,000,000 random samples of size 2. The program calculated ranges for each sample and computed $\bar{R}$ to be the average of the 1,000,000 sample ranges.

Figure 2 shows the individuals chart based on the control limits of $\bar{\bar{x}} \pm \frac{2}{3} E_2 \bar{R}$. Overall, the chart identifies only 1 out of 158 courses falling outside of the control limits. From the control chart, we conclude that all the other instructors perform well (at least within predictable limits) and only instructor 1's performance is unusually low.

Modified p Chart

In contrast, the modified *p* chart, shown in Figure 3, shows 10 out of 158 courses falling outside of the control limits, with 3 of them falling above the upper control limit (UCL) and 7

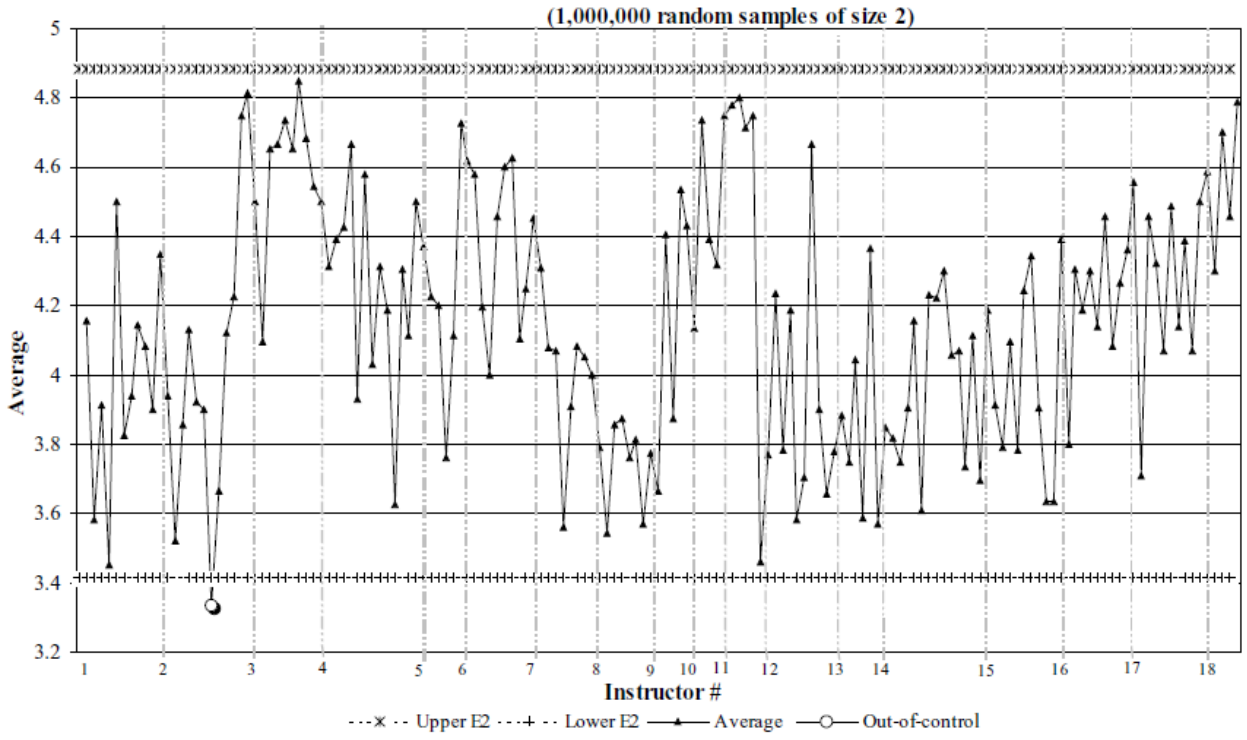

Figure 2: Single measurement chart.

below the lower control limit (LCL). Among the 18 instructors, instructors 1 and 3 are identified to perform considerably lower than average with 18.18% and 33.33% of course evaluations

falling below the LCL, respectively. By comparison, instructor 11 is identified to perform considerably higher than average with 40% of course evaluations located above the UCL.

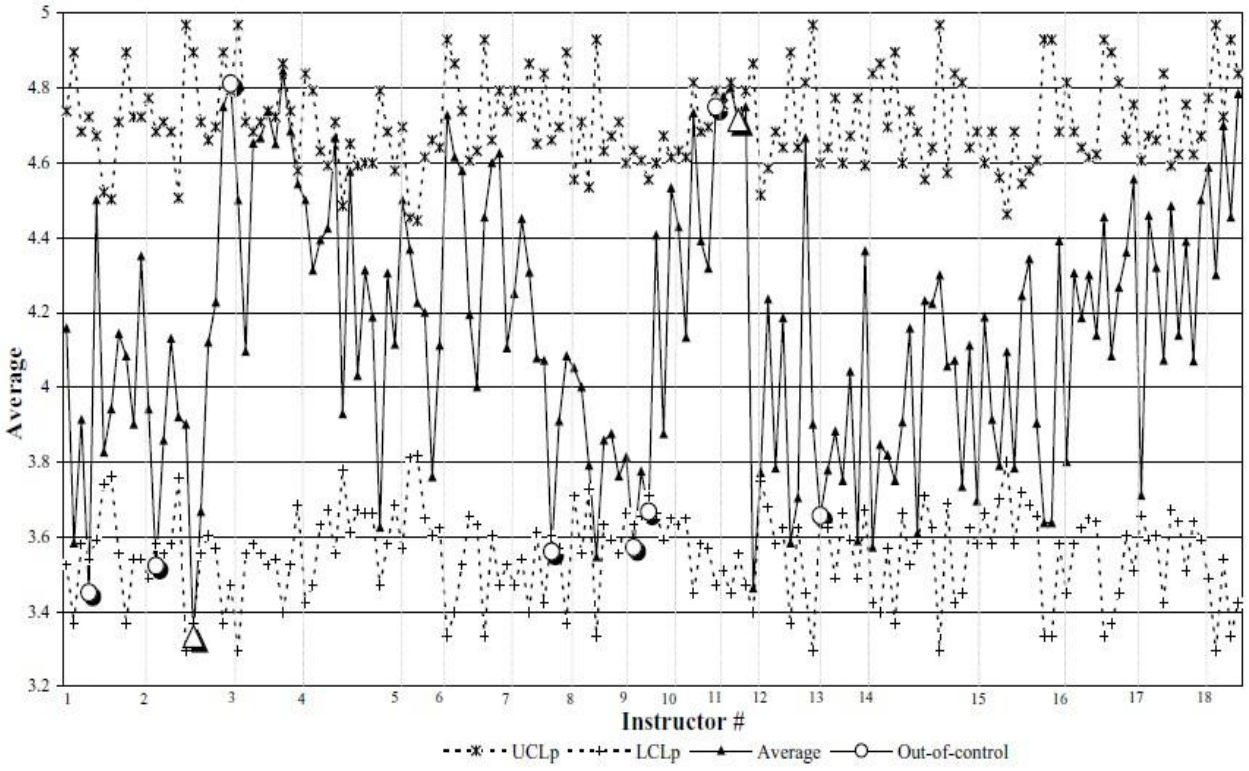

Figure 3: Modified *p* chart.

From the modified *p* chart, we identify five instructors (1, 2, 7, 9, and 13) that need to make improvement in their teaching performance by either developing a self-improvement plan or by consulting the other two instructors (3 and 11) whose performance is beyond the UCL.

z-Score Chart

The *z*-score chart (Figure 4) shows how far each course evaluation deviates from the average rating at the departmental level in terms of standard errors. Unlike the modified *p* chart, the *z*-score chart only identifies three instructors as being significantly different from the rest (i.e., three outside the control limits). They are instructors 2, 7, and 9, all of whom have scores below the lower limit. None of the instructors was shown to be significantly higher than the others, with no points falling above the upper limit.

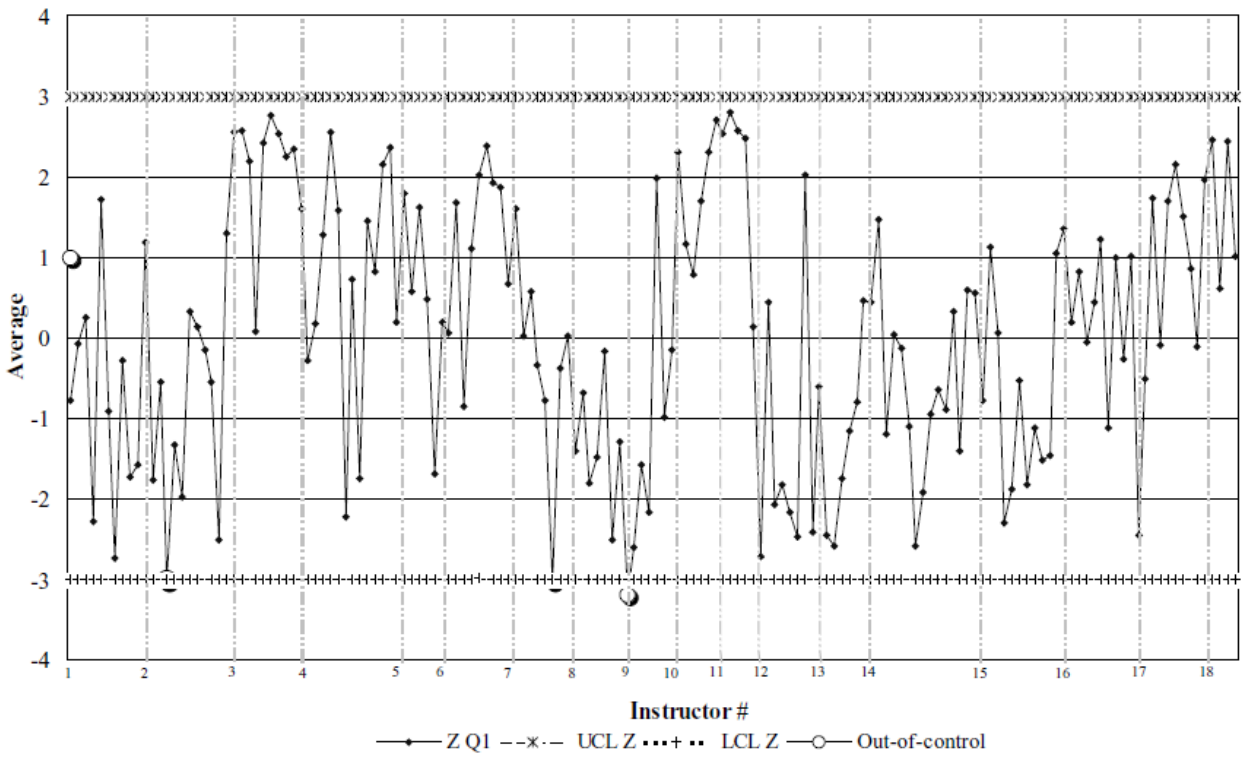

Figure 4: z-score chart.

Comparison and Further Discussion

The control charts (individuals, modified ⬚⬚, and ⬚⬚score charts) discussed in this article can individually identify sets of points falling out of the control limits and therefore provide us with insights regarding individual teaching performance and future improvement direction. Because they adopt different approaches in identifying points of interest, however, there is no consensus regarding which chart of teaching performance is the best. Therefore, we present a discussion of the relative merits and deficiencies of each approach. Table 1 provides a summary of the advantages and disadvantages associated with each of the three charts.

We start with the simplest measure to calculate, used in the ⬚⬚score chart. The primary advantage of this measure is its ease of computation. This measure only requires the average score of each course, plus the mean and standard deviation of each course at the departmental level. The calculated z scores present how far in terms of standard deviation each course evaluation deviates from the mean of the department. As a consequence, ⬚⬚score charts are useful tools for administrators to maintain a certain level of teaching performance. While easy to calculate and interpret, there are several drawbacks to ⬚⬚score charts. First of all, ⬚⬚ = ⬚⬚−μ ⬚⬚/√⬚⬚ which provides a higher estimate of the standard deviation, and hence wider relative control chart limits than those in the modified p chart. Second, there are no standard cutoff

values for what represents “good” agreement and no method of performing a statistical test if *3-sigma* limits are unable to identify any out-of-control points. Because of the drawback, the setting of appropriate specification limits for z-score charts is arbitrary and requires prior work and experience, which limits the application of z-score charts to SET data.

Sources	Single Measurement (Marks & Connell, 2003)	*z*-Score Chart	Modified *p* Chart
Advantages	• Intuitively logical • Can be used when only means are available to evaluators	• Easy to compute	• Applicable to nonnormally distributed SET data • Intuitively logical • Relatively easy to program into a computer
Disadvantages	• Subject to normal distribution • Do not have any time elements, invalidate the application of $\bar{x}$ chart in SET data	• Difficult to interpret and needs prior experience when *3-sigma* limits fail • Results in a higher estimate of the standard deviation and hence wider relative control limits	• Not well established in literature • Comparably complex in calculation

Table 1: Overview of methods for analyzing student evaluation of teaching (SET) data.

The individuals chart is applicable under circumstances where SET means must be considered as individual units that cannot be conveniently divided into subgroups. Because single measurements of variables constitute a subgroup of size 1, the estimate of the process variability is measured as the moving range (the absolute value of the difference between each data point and the one that immediately preceded it). The disadvantages for using single measurement charts include the possible correlation in the moving ranges as well as potentially inflated control limits. Marks and Connell’s approach of using random samples of size 2 does not include any time elements and therefore can only identify one course falling out of control in our sample, which leads to an overly optimistic attitude toward teaching performance. Finally, it ignores the fact that classes have different sizes—the means are essentially treated the same whether they were computed from a class of 100 students or one of 10.

The modified *p* chart is recommended in this article as the best approach to analyze SET data for several reasons. First, it is applicable to nonnormally distributed SET data. The modified *p* chart utilizes distributions that are appropriate for categorical data, and the control limits can

vary with the sample (class) size. Second, it is intuitively logical and only needs the computation of the proportion of students who gave response *x* among *k* possible responses. As demonstrated in Figure 3, the modified *p* chart identifies out-of-control points explicitly along the time horizon, which helps individual instructors and administrators monitor and improve teaching performance continuously. Compared with individual charts and *z*-score charts, the modified *p* chart (Wardell & Candia, 1996) is not well established in the literature and it involves comparably complex computations to calculate the control limits. However, it is relatively easy to program the modified *p* chart into a computer; therefore, it should be easy to calculate the limits in practice.

When we calculated and plotted the modified *p* chart, we ordered the courses along a time horizon for each instructor. However, there are some cases where one instructor taught multiple classes during a certain semester. Under such conditions, we can choose to either combine those multiple course evaluations into a single value or leave them as is. We did not combine the teaching evaluations into a single score because we think the composite or average score cannot be an accurate reflection of the actual teaching performance.

To justify our approach, we refer to the concept of rational subgrouping proposed by Shewhart, where rational subgroups are composed of items which were produced under essentially the same conditions and each subgroup is formed by using consecutive units. In this article, all the evaluations are subgrouped by individual instructors and the evaluations within the same subgroup reflect the teaching performance of the same instructor.

A rational subgroup is simply "a sample in which all of the items are produced under conditions in which only random effects are responsible for the observed variation" (Nelson, 1988). The rational subgroup has the following characteristics:

- The observations composing the subgroup are independent.
- The observations within a subgroup are from a single, stable process.
- The subgroups are formed from observations taken in a time-ordered sequence.

The observations composing the subgroup in our study are independent and are from a single, stable process. In addition, the observations forming the subgroup are arranged in a time-ordered sequence. In case one instructor taught multiple classes in a certain semester, the teaching evaluations for that instructor still meet the criteria because they are arranged in a time order considering all the rest of the courses taught by that instructor in other semesters.

To demonstrate our preference for the modified *p* chart over the other charts, we take the teaching evaluation scores for instructors 2 and 11 as examples. For instructor 2, the modified *p* chart identifies the eighth course (course 798, Autumn, 1996, presented by the first triangle in Figure 3) as falling below the LCL. None of the other charts identified this course as unusual. While comparing the out-of-control course with other classes taught by the same instructor, we find it is the first time for instructor 2 to teach course 798. In addition, the

instructor taught three courses for the same semester, which is unusual during our observation period. Based on the above observation, it is reasonable to expect the instructor to perform slightly lower than average for the class.

We next consider instructor 11, who performed extremely well during our observation period. On average, around 80% of the students rated the instructor at 5 and no students rated her/him below 3. Initially, we had the impression that instructor 11 performed far beyond other instructors in the business school and there is no need for him/her to make any further improvement, which is what we saw the individuals and *z*-score charts indicate. We still see variation in the performance for instructor 11, however, the modified *p* chart identified two courses outside of the UCLs, which are courses 520, Spring semesters of 1995 and 1996 (presented by the second triangle in Figure 3).

As discussed above, to understand more completely the difference among the three charts discussed in this article, we have developed a list of several advantages and disadvantages for the three methods of monitoring SET data (see Table 1). Obviously, there are trade-offs between the charts, and each researcher or practitioner must make his or her own decision regarding which method is most appropriate. On the basis of our analysis and comparison, we suggest the modified *p* chart as the best alternative for analyzing SET data.

Conclusion

We have provided a comparison of methods for analyzing SET data over time. Traditional ��charts that assume normally distributed data may be biased because SET data are categorical and hence not normal. In addition, the large and varying sample sizes make the implementation of traditional charts complicated. To address this difficulty, we proposed two charts, namely, a modified p chart and a ��score chart, and compared them to an individuals chart proposed by Marks and Connell (2003) using teaching evaluation data from an eastern U.S. university. In spite of the relatively complex computation of control limits, the modified *p* chart is shown to have better performance in terms of explicitly identifying out-of-control points and utilizing the actual distribution of SET data than the other two charts.

This article presents the research opportunity in resolving the issues surrounding the use of control charts to monitor SET data. One of the issues is to determine the applicability of various control charts given the information contained in the database. First, we conclude the modified *p* chart is suitable to the nonnormally distributed data set. Second, under certain extreme situations where administrators are only presented with summary means, individual charts can be used to monitor evaluation scores. Other issues include making identification of special causes for the out-of-control points, generating recommendations for underperforming instructors, and automatically reporting on the teaching performance for each instructor and administrator. The above issues raise questions for further research in both quality control and decision support areas.

References

Brightman, H. (1987). Towards teaching excellence in the decision sciences. *Decision Sciences, 18*, 646–661.

Brightman, H., Elliott, M., & Bhada, Y. (1993). Increasing the effectiveness of student evaluation of instructor data through a factor score comparative report. *Decision Sciences, 24*(1), 192–199.

Cashin, W. (1995). Student ratings of teaching: The research revisited. Idea Paper No. 32, Center for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University.

Centra, A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

Davis, B. (1995). Measure process capability when *Cp* won't work. *Quality, 34*(8), 20–21.

Evans, J. (1986). Creative thinking and innovation education in the decision sciences. *Decision Sciences, 17*, 250–262.

Evans, J., & Lindsay, W. (1999). *The management and control of quality*. Cincinnati, OH: South-Western College Publishing.

Feldman, K. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education, 6*, 233–274.

Gitlow, H., Gitlow, S., Oppenheim, A., & Oppenheim, R. (1995). *Quality management: Tools and methods for improvement*. Homewood, IL: Richard D. Irwin, Inc.

Herbert, M., & Lawrence, R. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197.

Kanagaretnam, K., Mathieu, R., & Thevaranjan, A. (2003). An economic analysis of the use of student evaluations: Implications for universities. *Managerial and Decisional Economics, 24*(1), 1–13.

Katz, D. (1973). Faculty salaries, promotions, and productivity at a large university. *American Economic Review, 63*, 469–477.

Marks, N., & Connell, R. (2003). Using statistical control charts to analyze data from student evaluations of teaching. *Decision Sciences Journal of Innovative Education, 1*(2), 259–272.

Marsh, H., & Dillon, K. (1980). Academic productivity and faculty supplemental income. *Journal of Higher Education, 51*, 546–555.

Martin, J. (1998). Evaluating faculty based on student opinions: Problems, implications and recommendations from Deming's theory of management perspective. *Issues in Accounting Education, 13*(4), 1079–1094.

Mesak, H., & Jauch, L. (1991). Faculty performance evaluation: Modeling to Improve personnel decisions. *Decision Sciences, 22*(5), 1142–1157.

McKeachie, J. (1979). Student ratings of faculty: A reprise. *Academe, 65*, 384–397.

Nelson, L. (1988). Control charts: Rational subgroups and effective applications. *Journal of Quality Technology, 20*(1), 73–75.

Ozgur, C., & Strasser, S. (2004). A study of the statistical inference criteria: Can we agree on when to use Z versus t? *Decision Sciences Journal of lnnovative Education, 2*(2), 177–192.

Ronald, S. (1988). Maximize Z scores and outliers. *American Statistician, 42*(1), 79–81.

Siegfried, J., & White, K. (1973). Financial rewards to research and teaching: A case study of academic economists author. *American Economic Review, 63*, 309–315.

Sincich, T. (1986). *Business statistics by examples* (2nd ed.). San Francisco: Dellen.

Stratton, R., Steven, M., & Randall, K. (1994). Faculty behavior, grades, and student evaluations. *Journal of Economic Education, 25*(1), 5–15.

Sylvia, A., & Phillip, A. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198–1208.

Wallace, J., &Wallace, W. (1998). Why the costs of student evaluations have long since exceeded their value. *Issues in Accounting Education, 13*(2), 443–447.

Wardell, D., & Candia, M. (1996). Statistical process monitoring of customer satisfaction survey data. *Quality Management Journal, 3*(4), 36–50.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *Chronicle of Higher Education, 44*(19), 12–14.

Younger, M. (1979). *A handbook for linear regression*. North Scituate, MA: Duxbury Press.

Author Bios

Xin (David) Ding is a PhD candidate at the David Eccles School of Business at University of Utah. He received his MS in engineering from the South China University of Technology. His research interests include service quality, quality control, service operations, and e-commerce. He has contributed to proceedings and presentations at POMS, DSI, and INFORMS annual conferences. He is a member of the Decision Science Institute, Production and Operations Management Society, Institute for Operations Research, and the Management Sciences.

Don G. Wardell is Associate Professor and Chair of the Department of Management at the University of Utah's David Eccles School of Business (DESB). He received BS and MS degrees in Metallurgical Engineering from the University of Utah and a PhD degree from Purdue University's Krannert Graduate School of Management. Dr. Wardell has taught at both the undergraduate and graduate levels, including teaching classes in Spanish at INCAE in Costa Rica. Dr. Wardell was honored with the DESB's Masters Teaching Excellence Award, the Brady Superior Teaching Award, and the Marvin J. Ashton Award for Excellence in Undergraduate Teaching. His research interests are mainly in the area of quality management, especially statistical process control. He has served as an associate editor for *Technometrics*, is a member of the editorial review boards of *Production and Operations Management* and *IIE Transactions on Quality and Reliability,* and reviews articles for numerous journals.

Rohit Verma is Associate Professor of Operations Management and Thayne Robson Fellow at the David Eccles School of Business, University of Utah. His research interests include Product/Service Design, Operations Strategy, Process Improvement, and Operations/Marketing interrelated issues. His research has appeared in *Decision Sciences, Journal of Operations Management, Journal of Product Innovation Management, Journal of Service Research, Omega, Production and Operations Management,* and other journals. Rohit received the Skinner Award for Early Career Research Accomplishments from the Production and Operations Management Society in April 2001. He is also one of the first recipients of the Sprit of Inquiry Award, the highest honor for scholarly activities within DePaul University where he taught from 1995 to 2001.