Statistical Process Control Charts for Measuring Rating Consistency Over Time

M.H. Omar
Dept of Mathematics and Statistics
King Fahd University of Petroleum and Minerals
Dhahran, 31261 Saudi Arabia.
Email: omarmh@kfupm.edu.sa

Abstract

Methods of Statistical Process Control (SPC) have been briefly explored in the field of educational measurement as early as 1999. However, only the use of CUSUM chart was explored. In this paper, other methods of statistical quality control is introduced and explored. In particular, an SPC method of Shewhart control charts is introduced as a potential technique in ensuring quality in a statistical process of rating performance assessments. Several strengths and weaknesses of the procedure are explored with illustrative simulated rating data. Further research directions are also suggested.

Introduction

Statistical Process Control charts have been widely used in the industrial setting as backbone tools for maintaining product quality. W. A. Shewhart and W. E. Deming were two well-known statisticians responsible for introducing and popularizing these in-stream statistical mechanisms for controlling the manufacturing process. W. A. Shewhart was very instrumental in introducing what is known as the Shewhart control chart. The main use of this chart is to track the means of batches of industrial products so that they may be produced within industrial specifications. These in-stream process control ideas were popularized by W. E. Deming's Total Quality Management (TQM) idea and were used extensively by Japanese industries to produce high quality products (Deming, 1981). The Japanese found that technicians with close to no training in statistics can track quality of products fairly well because all they have to do is to graphically check if the means randomly fall in a certain band of tolerance to signify acceptability of the product produced by the process.

Although the business industries have for decades extensively used these statistical process control charts to ensure quality, their use in the educational measurement fields have been quite limited. Thus, the main purpose of this paper is to explore how these tools can be adopted for use as quality maintaining tools for an educational measurement process.

Literature

What is Statistical Process Control?

Statistical Process Control (SPC) is a control mechanism whereby measurements of product quality are actively obtained and charted simultaneously as industrial products are produced. Control is obtained when a statistical measurement such as means of a group of products are within certain control limits drawn on the statistical process chart. For these charts, there are certain set of rules to follow that will tell the technicians when a process may be out of control. When these conditions are observed, the technicians are expected to stop the manufacturing process so that corrective actions can be taken.

Rational Subgroups

Taking measurements of products on an assembly line can be costly or can potentially damage product quality. Naturally, companies often do not measure all products produced on an assembly line as this may prove to be counter-productive or economically not feasible. A compromise is often taken. For measurement of consistent product quality, most companies select what is termed as 'rational subgroups' of products. Some random sampling scheme is usually employed to secure these 'rational subgroups'. These 'rational subgroups' are then laboriously checked for quality. Measurements emanating from these 'rational subgroups' are then charted for quality control purposes.

Rules for out of control process

For the SPC charts, a process is considered out of control or drifting out of control when one or more of the following is observed (Montgomery, 2005):

1. Any point is outside the upper or lower 3 sigma control limit.
2. At least eight consecutive points are on only one side of the control chart.
3. Two or three consecutive points are outside the upper or lower 2 sigma warning limit but inside the control limits.
4. Four or five consecutive points beyond the 1-sigma limits.
5. An unusual or random pattern in the data (such as a cyclic pattern).
6. One or more points near a warning or control limit.

SPC applications in Educational Measurement to date

Use of cumulative sum control chart for identifying a person-fit index in a Computer Adaptive Testing (CAT) environment has been introduced by Krimpen-Stoop, Edith, and Meijer

(1999) and Edith, Krimpen-Stoop, and Meijer (2001). Meijer (2002) also used the CUSUM-based person-fit index to detect outliers in a high-stakes certification testing. Veerkamp and Glas (2000) used CUSUM charts to detect drifts in 1PL and 3PL item parameter values in a CAT environment.

Beyond these, no other use of statistical process control has been explored in the field of educational measurement. This leads one to ask why? In particular, why do these quality ensuring techniques lend themselves very easily to measure production of industrial products but are still quite foreign to the field of educational measurement. One may venture at least two reasons. First, industrial products are mass produced to the same specifications by the same or similar machine. Thus automation of production is commonplace and data gathering is not complicated by multifaceted specifications. Second, with exception to some CAT environments, multiple-choice data are machine scored cross-sectionally while statistical process control charts are useful tools when data are collected in a repetitive or time-sequenced manner such as with mass production of products.

Thus, with these reasons we rarely see these quality control techniques applied in the educational measurement field. As the name "statistical process control" involves control of a process, beyond person fit in the CAT environment, another unexplored area of educational measurement that may involve a process needing control is ratings of performance assessments. The ratings of papers involves a process of implementing a set of rating standards or scoring rubric to a set of student papers such as written mathematics problem solving papers or essays.

A potential application in Educational Measurement: Monitoring Rating Consistency Over Time

Technical aspects of rating performance assessment to date include, among others, inter-rater reliability index which requires the same set of papers rated twice by two separate raters. The correlation between ratings from these two raters is considered as a source of rating reliability. Usually this type of reliability evidence is gathered at the onset of the rating process and is mainly indicative of consistent implementation of the scoring rubric by two or more independent raters. However, inter-rater reliability index does not guarantee that raters with consistent rubric implementation do rate stably over time. This is where the process could be

monitored with a statistical process control (SPC) technique at specific intervals in the rating process.

Method

Because no performance assessment data to date is monitored with SPC techniques, the behaviour of performance ratings in this paper is studied with the help of simulated data using MINITAB version 13.

Scoring Rubric

A typical scoring rubric for a mathematics Problem Solving question may involve a 5-point rubric system. A 5 point is obtained when a paper is fully correct with complete steps and interpretation. A 4 point may be obtained when a paper is fully correct with complete steps except interpretation. A 3 point may contain one incomplete step with no correct interpretation. A 2 point may include the correct problem set-up and a correct first step but other steps are either incomplete or incorrect. A 1 point paper may involve a correct problem set-up but all steps are incorrect. While a 0 point paper contains no work or work totally unrelated to the question. Solution steps are often provided as part of the rubric to guide raters while rating student papers. This detailed rubric is usually applied in a range-finding activity and tested and implemented by raters on a set of anchor papers in a rater training.

Monitored Papers

For illustration purposes, a set of 5 papers were randomly chosen at each monitoring period from the past rating periods. As such, each of these 5 monitored papers would already have a score associated with it. These monitored papers were rated again at specific intervals in the rating process. Several strategies can be offered to avoid rater's memory effects on papers that have been previously rated. These include:

1. Random shuffling of monitored papers.
2. Exact replicas of papers except with different handwriting.
3. A typed up version of papers.
4. A combination of above.

Readers can also imagine other strategies that would achieve the same effects.

Rating Scheme with Statistical Process Control

To periodically monitor the rating process, the following steps should be followed:

1. A set of five or more previously rated papers are randomly selected as monitored papers. The rater rates these 5 papers and descriptive statistics are collected.
2. The initial ratings for these papers are subtracted from the second ratings. The mean of the difference in ratings are graphed.
3. The rater continues to rate other batches of 5 papers and rating scores for each paper and their summary statistics are recorded.
4. When a specific number, say 60, of papers is rated, another set of 5 previously rated papers are rated again and descriptive statistics of the rating difference are again collected and plotted in the same chart in step (2).
5. Repeat steps (3) and (4) and monitor the charted means of difference in ratings for whether it violates either control or warning limits and whether any sign of nonrandomness are observed.

The ratings and re-ratings of these monitored papers are captured and charted to maintain quality control of the rating standards and to detect any signs of rating drifts or ratings that are out of control. If there is no difference in ratings, one would tend to observe predominantly 0 mean differences. Otherwise, substantially larger or lower effects will be observed.

Simulated Data

The simulated data for this study is obtained for 5 raters over 25 monitoring periods. For each monitoring period, each rater would rate a set of papers he/she had rated before except that the second set of paper is a re-copied version of the original. In reality, this can be accomplished in various ways as discussed in the monitored papers section. Then, the second set of ratings is subtracted from the first. Due to a previously completed rater training or range-finding activity, a standard deviation of rating differences on anchor papers are known and can be used as population control values. For this paper, the population standard deviation was 0.75.

To illustrate the behaviour of rating processes that are in control and those that are out of control, the following provides description of the parameters of the sets of 5 ratings for each rater at each of the 25 monitoring periods:

1. Rater 1 is in control.
   The observed rating for the monitored papers is sampled from a normal population with mean of 3 and standard deviation of 1. The observed difference between the second and the first rating for the same monitored papers is sampled from a normal population with mean difference of 0 and standard deviation of differences of 0.65.
2. Rater 2 violates the upper control limit.
   Observed mean ratings are sampled like for rater 1 except a large disturbance is added to the mean difference at monitoring periods 18 and 19. The disturbance is normally distributed with mean of 1.3 and standard deviation of 0.5.

3. Rater 3 violates the lower warning limit.
   Observed mean ratings are sampled like for rater 1 except a big disturbance is subtracted from the mean difference between monitoring periods 14 and 16 inclusive. The disturbance is normally distributed with mean of 0.9 and standard deviation of 0.5.
4. Rater 4 shows signs of nonrandomness.
   Observed mean ratings are sampled like for rater 1 except more negative runs are sampled towards the end of the monitoring process. That is, a medium sized disturbance is subtracted from the mean difference between monitoring periods 14 and 16 inclusive. The disturbance is normally distributed with mean of 0.55 and standard deviation of 0.5.
5. Rater 5 shows signs of out of control monitoring period variation with medium size disturbance.
   Observed mean ratings are sampled like for rater 1 except between monitoring periods 2 and 5 inclusive, there is an addition of a medium-sized disturbance with large variability. The disturbance is normally distributed with mean of 0.4 and standard deviation of 1.5.

How to Build a Shewhart SPC Control Chart

To build the Shewhart control chart, the standard deviation of the initial rating difference (such as those resulting from a rater training) can be used as an estimate of the population standard deviation. With this, the following steps are followed:

1. The mean rating difference of 0 can be used as the center or target. $(\mu_D = 0)$
2. Standard deviation of the rating difference from the rater training multiplied by 2 and divided by square root of monitored paper size is added and subtracted from the target to obtain the upper and lower warning limits. $(\mu_D \pm 2\sigma_D / \sqrt{n})$
3. Standard deviation in (2) multiplied by 3 and divided by square root of monitored paper size is added and subtracted from the target to obtain the upper and lower control limits. $(\mu_D \pm 3\sigma_D / \sqrt{n})$

For example, the upper control limit is $0 + 3(0.75)/\sqrt{5} = 1.00623.$

How to Build an S Chart

Often it may not suffice to monitor means of rating difference only. An S chart can also be produced to see if variabilities of rating differences in the monitored papers are fairly homogeneous between monitoring periods. To build the S chart, the following steps are followed (see Montgomery, 1991).

1. The standard deviation of the rating difference from the rater training is used as the center or target of the control chart. $(c_4\sigma)$

2. The lower control limit is obtained by $(c_4 - 3\sqrt{1-c_4^2})\sigma$
3. The upper control limit is obtained by $(c_4 + 3\sqrt{1-c_4^2})\sigma$
4. The lower warning limit is obtained by $(c_4 - 2\sqrt{1-c_4^2})\sigma$ and
5. The upper warning limit is obtained from $(c_4 + 2\sqrt{1-c_4^2})\sigma$

where

$c_4 = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$. For n = 5 in this article, $c_4 = \frac{3}{8}\sqrt{2\pi} = 0.939986.$

Montgomery (1991, p. 232) commented that $E[S] = c_4\sigma$. It is also important to note that if any of the lower limits are negative, a limit value of 0 is used instead since the population standard deviations cannot be negative.

Results and Discussions

Figure 1 provides a Shewhart control chart of the mean rating differences by a single rater over several quality control periods. As explained previously, it is not difficult to see that since no points are beyond the warning limits and the pattern is preety much random, the mean rating differences for the monitored papers are on target and are stable over time. Figure 2, 3, and 4 provide similar Shewhart SPC control charts on the same papers but by different raters. Figure 2 shows signs of flagrant violation of the Shewhart SPC chart for monitoring periods 18 to 19 as these points are above the 3-sigma upper control limits. Figure 3 shows violations of the 2-sigma lower warning limit for periods 14 to 16. Finally, Figure 4 shows signs of nonrandomness from periods 16 to 25 as the mean ratings are consistently on the lower side of the chart.

Insert Figure 1 about here

Insert Figure 2 about here

Insert Figure 3 about here

Insert Figure 4 about here

Figure 5 shows an S chart for ratings by rater 5. Periods 2 to 5 shows larger variation than expected as these standard deviations are outside the 3-sigma upper control limit of an S chart.

It can clearly be seen that with such charts, it is possible to communicate to each rater the quality of his/her ratings from time to time. What's also important is that with current computer technology, such charts can be simultaneously produced while raters rate papers. As such, raters can be very much aware of their level of rating consistency over time. It is also very possible to stop mass scoring for corrective actions when some indications of 'out-of-control' ratings are observed.

Thus, it seems that the SPC charts hold some potential and promise if used simultaneously to monitor stability of performance ratings over time.

Limitations and Further Directions

Although quality control of randomly chosen monitored papers can be accomplished as outlined in this paper, most papers are not rated again. Thus, like mass production in industrial and business settings, there is a small possibility that some aberrant ratings would creep into the mass scoring. However, with such initiatives for periodic monitoring of scored papers, the possibility of 'out-of-control' rating drifts is probably much less.

In this article, the Shewhart chart has a control value of 0 mean differences while in the S chart a target value of σ was used as they were simulated to come from a previous rater training. However in some situations, if these target values are not available and must be estimated, slight modifications of the control charts are necessary.

In the Methods section of this article, the recommendation of 5 monitored papers in step (1) and 60 papers between successive monitoring periods in step (4) are currently only as good as any other recommendation. Since this is one of the first research on implementation of SPC technique in the area of performance assessment, no empirical backing on the number 60 or the 5 monitored papers has yet been produced. The author urges other researchers to pursue the best empirical number between successive re-ratings of monitored papers under (1) economic and (2) pragmatic concerns.

Reference

Edith, M. L. A., Krimpen-Stoop, V., & Meijer, R. R. (2001). CUSUM-Based Person-Fit Statistics for Adaptive Testing**,** *Journal of Educational and Behavioral Statistics*, 26(2), 199-217.

Krimpen-Stoop, V., Edith, M. L. A., & Meijer, R. R. (1999). CUSUM-Based Person-Fit Statistics for Adaptive Testing. *Research Report 99-05*, University of Twente, Enschede, The Netherlands.

Meijer, R. R. (2002). Outlier Detection in High-Stakes Certification Testing. Journal of Educational Measurement, 39(3), 219–233.

Veerkamp, W. J. J. & Glas, C. A.W. (2000). Detection of Known Items in Adaptive Testing with a Statistical Quality Control Method. *Journal of Educational and Behavioral Statistics*, 25(4), 373-389.

Deming, W.E. (1981). *Japanese Methods for Productivity and Quality*. George Washington University Press, Washington DC.

Montgomery, D.C. (1991). Introduction to Statistical Quality Control (2nd Ed), John Wiley & Sons, Singapore.

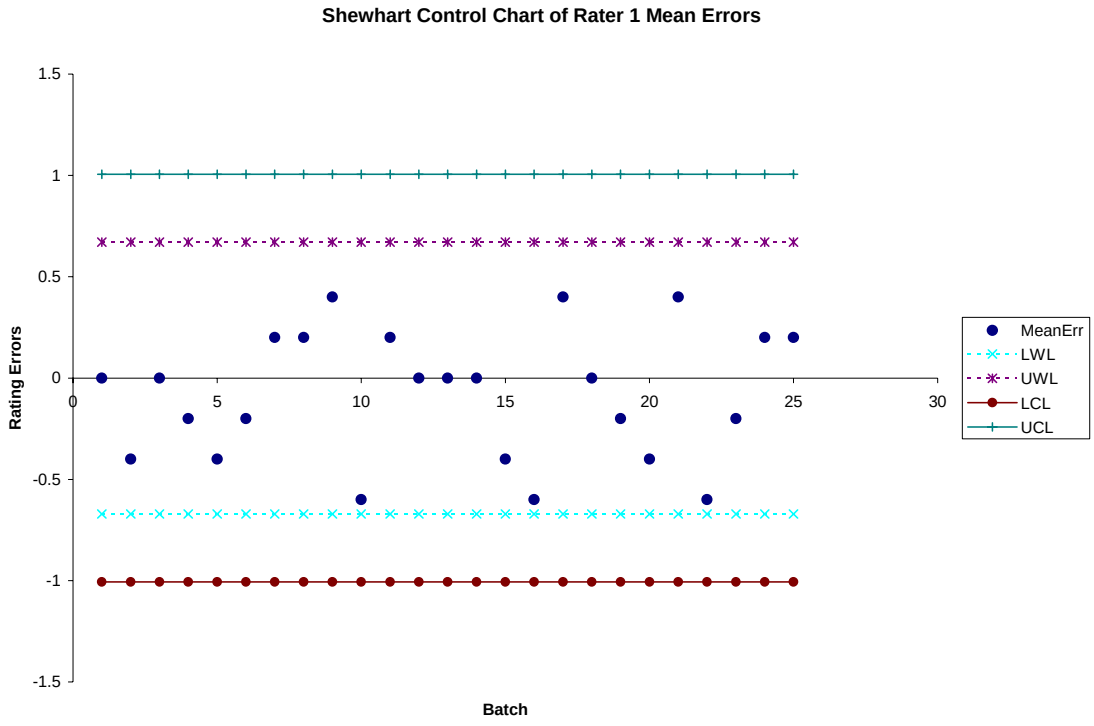

Figure 1. Rater 1 Mean Rating Differences across several Monitoring Periods.

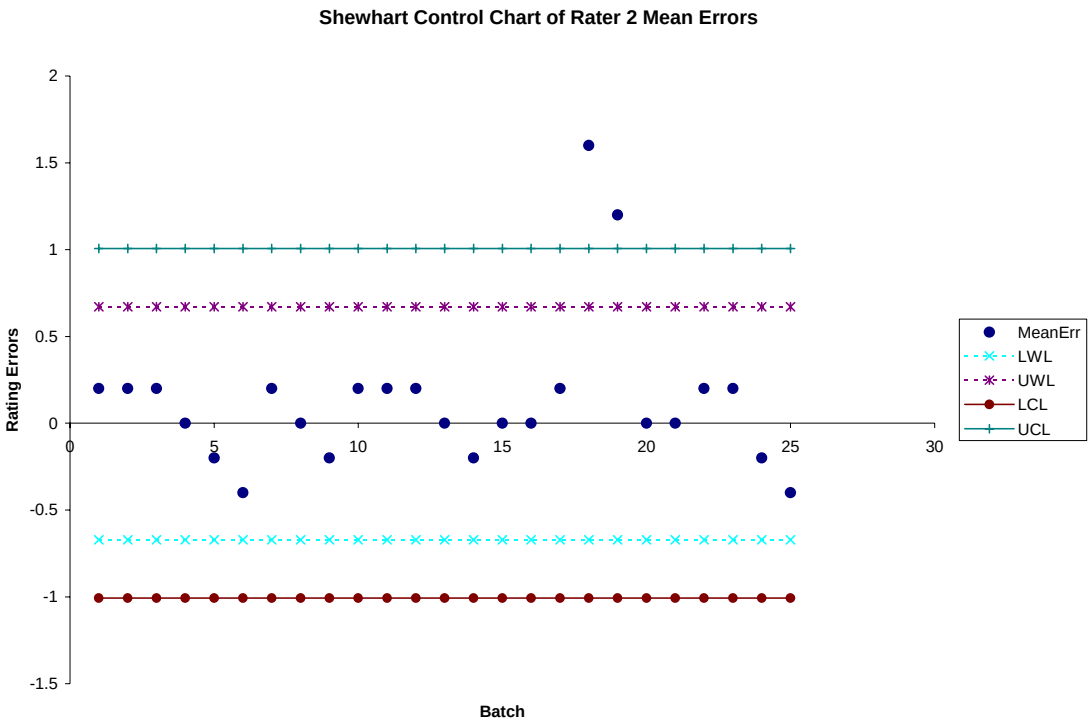

Figure 2. Rater 2 Mean Rating Differences across several Monitoring Periods.

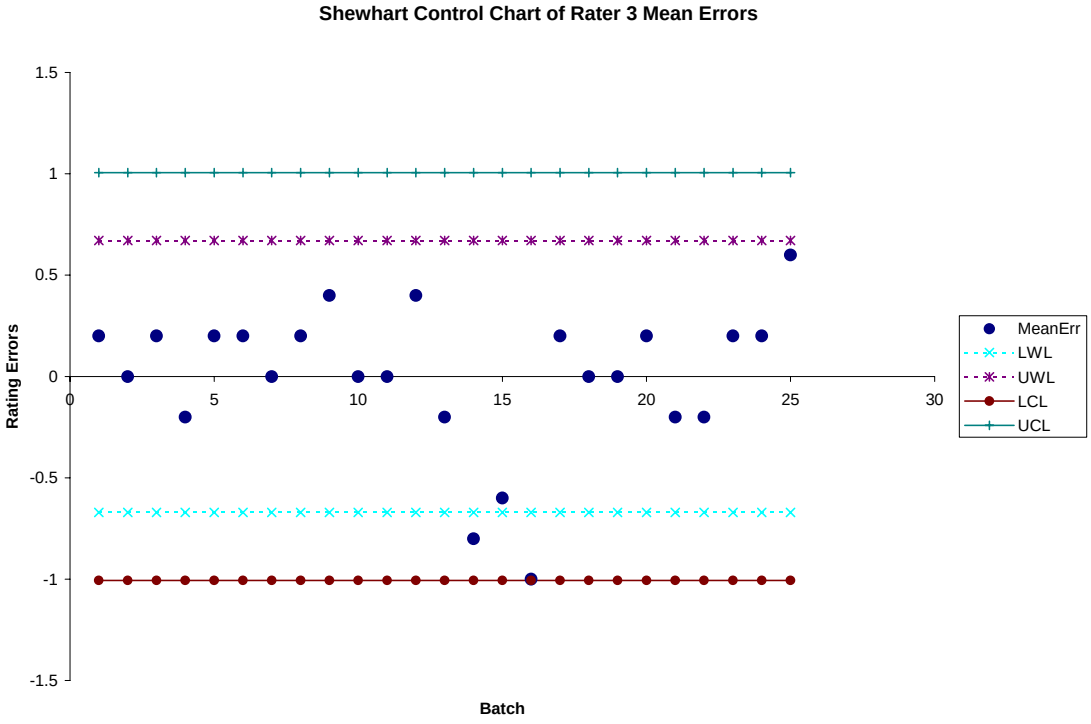

Figure 3. Rater 3 Mean Rating Differences across several Monitoring Periods.

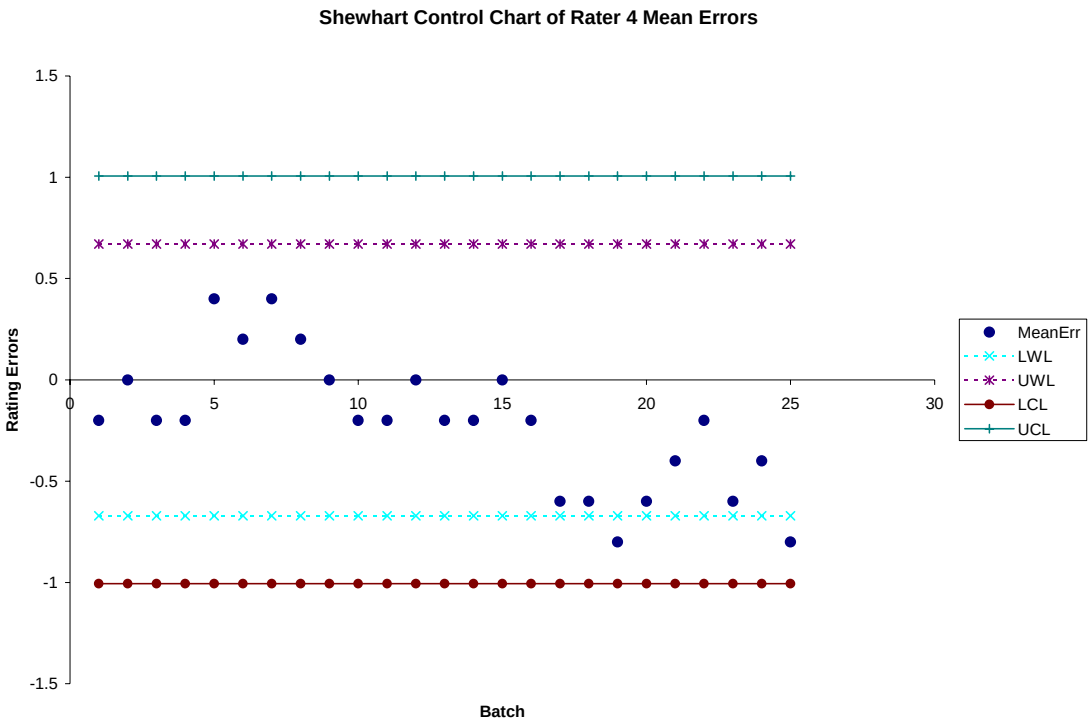

Figure 4. Rater 4 Mean Rating Differences across several Monitoring Periods.

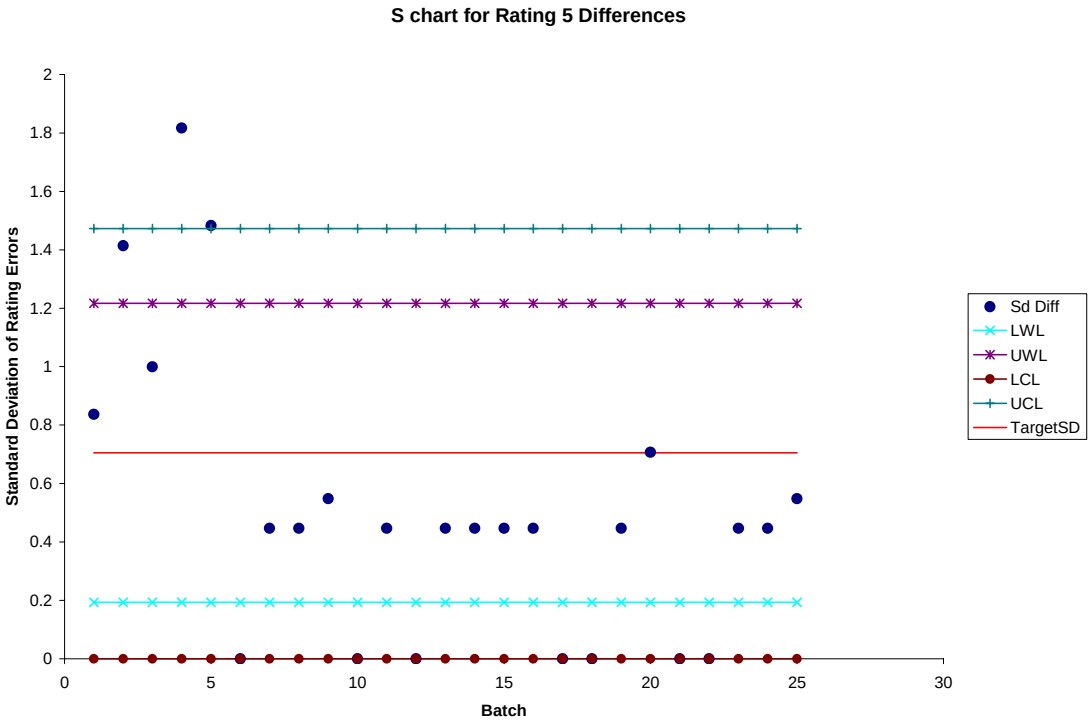

Figure 5. Rater 5 Mean Rating Differences across several Monitoring Periods.