Statistical Process Control Charts for Measuring and Monitoring Temporal Consistency of Ratings

M. Hafidz Omar
King Fahd University of Petroleum and Minerals

Methods of statistical process control were briefly investigated in the field of educational measurement as early as 1999. However, only the use of a cumulative sum chart was explored. In this article other methods of statistical quality control are introduced and explored. In particular, methods in the form of Shewhart mean and standard deviation charts are introduced as techniques for ensuring quality in a measurement process for rating performance items in operational assessments. Several strengths and weaknesses of the procedures are explored with illustrative real and simulated rating data. Further research directions are also suggested.

Quality assurance methods typically collect process evidence after a production process has been completed. In contrast, statistical process control (SPC) is a dynamic monitoring method where product quality is actively measured and simultaneously charted while manufactured goods are being mass produced. Statistical control is achieved when an index such as the means of groups of products are plotted within certain limits drawn on a chart.

In addition to consistency, production efficiency is important in mass production. Thus, a production line may not be stopped or slowed down unless there are indications of process shifts as indicated by the SPC charts. For these charts, rules exist to tell the technicians if a process is potentially out of statistical control. When these conditions exist the technicians are expected to stop the manufacturing process for corrective action.

Although SPC charts are meant for monitoring a production process as a whole, taking product measurements on an assembly line can be costly, can slow down a process, or can diminish product quality. Naturally, companies do not usually measure all products as this may be counterproductive or economically unfeasible. As a compromise, most companies select rational subgroups of products often using a random sampling scheme. These subgroups are then laboriously checked for quality and their summary measurements are charted for quality control purposes.

The use of SPC charts for quality control can be traced back to the earlier part of the twentieth century. The Shewhart (1931) control chart is used to track means of batches of industrial products so that these products may be manufactured within industrial specifications. These in-stream process control procedures were later popularized by Deming’s (1951) total quality management (TQM) and have been used extensively by Japanese industries to produce high-quality products. The Japanese have found that technicians with little to no training in statistics can track product quality fairly well because they only need to visually check if batch means randomly fall in a certain band of process tolerance on a graph (Deming, 1967, 1981). Since
then, these SPC charts have been widely used in industrial settings as tools for maintaining product quality.

Presently, the American Society for Quality Control (ASQC) includes SPC charts as an important tested domain to professionally certify process engineers (Vani, 1995). The main purpose of this article is to explore how they can be used further in educational measurement.

**Literature**

Although SPC techniques have been widespread in industrial settings, in the educational measurement field these techniques have received attention only recently. Van Krimpen-Stoop and Meijer (1999, 2001) have used cumulative sum (CUSUM) control charts to develop a person-fit index in a computer adaptive testing (CAT) environment. Meijer (2002) has also applied the CUSUM-based person-fit index to detect outliers in high-stakes certification testing. Veerkamp and Glas (2000) have detected drifts in 1PL and 3PL item parameter estimates in a CAT environment using CUSUM charts. All these studies either used one type of SPC technique (CUSUM) or are specialized for IRT-based applications. Beyond these, no other use of SPC charts have been explored in the field. Montgomery (1991, p. 258) has presented a partial reason by observing that “the product is different” has been a common misconception hindering use of SPC charts in nonmanufacturing applications. Because specification limits have rarely been required for products in nonmanufacturing environments, the process targets and capability have typically been undefined. Thus more imagination has been required to select the proper variable(s) for measurement. Montgomery has, however, commented that if measurements reflective of product quality, function, or performance could be accomplished, then the product’s nature has no bearing on the SPC charts’ general applicability. According to Nelson (1988), SPC charts have also been applied to other nonmanufacturing situations such as tracking stability of engineering measurement processes, examining errors in forecasting, and economic modeling.

**Statistical Basis of the Control Charts**

On the basis of probability as defined by Tshebycheff’s inequality, Shewhart (1931) originally chose three standard errors (3-sigma units) as the marker for data points exceeding the limits of statistical control. This inequality states that no less than 89% of a sampling distribution, no matter the shape, lies within three standard errors of its mean. In general, \( P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - 1/k^2 \) for every \( k > 0 \). If the underlying distribution is normal and the process is in control, 99.73% of the monitored means lie within these 3-sigma control limits. On the other hand, if the statistics to be monitored were not normal, at worst, 89% of these would be within the control limits. With a more refined version of Tshebycheff’s inequality (Camp, 1922; Does, Roes, & Trip, 1991; Shewhart, 1931) for strongly unimodal statistical distributions (e.g., that of the standard deviation estimator, \( s \)), at least 95% of the monitored statistics would be within the limits. For the Shewhart mean (\( \bar{X} \)) chart, by virtue of the central limit theorem, normality further increases the percentage of means within 3-sigma control limits to around 99.73%. Some researchers
(Burr, 1967; Schilling & Nelson, 1976) have found nonnormality effects on Shewhart control charts to be minimal.

Historically, the choice of 3-sigma for control limits has been based on providing a proper balance between control for Type I and Type II errors. For the manufacturing process, Shewhart also found the coefficient 3 to be excellent in providing this balance (Nelson, 1988). As a result, the American Society for Testing and Materials (ASTM, 1976, pp. 134–135) printed tables for control limits using 3-sigma for $\bar{X}$ and standard deviation ($s$) charts. A joint reference manual for SPC charts (American Society for Quality Control—Automotive Division & Automotive Industry Action Group of GM, Ford, & Chrysler, 1991) also uses 3-sigma limits. Although the manufacturing industry has widely adopted these 3-sigma limits for their processes (Montgomery, 1991, p. 217), other coefficients may be adopted for a particular application by carefully balancing the effects of Type I and II errors (e.g., Hillier, 1969).

**A Potential SPC Application: Monitoring Rating Consistency over Time**

Beyond person fit in the CAT environment, an unexplored area involving a process that could use statistical control is scoring performance assessments. Rating consistency is particularly important when papers are scored in mass quantities.

The application of SPC techniques in scoring performance assessments, as pointed out by Montgomery, only requires identifying a variable that represents quality for the process needing statistical control. Rater reliability of open-ended scoring to date has been estimated with an interrater index that requires the same papers to be rated independently by at least two raters, and by computing percentages of exact agreement and agreement within one score point. For example, the Minnesota Comprehensive Assessment (MCA) II uses interrater correlation as a quality control index (Minnesota Department of Education [MDE] & Pearson Educational Measurement [PEM], 2006). For the MCA II, the interrater reliability index is collected daily during rater training and is often supplemented by cross tabulations of item scores by raters. When an unresolved rating inconsistency occurs, a quality control step of bringing a third rater to provide a final rating is frequently used. Many state assessment programs such as the Illinois Standard Achievement Test (ISAT) use this method to resolve large rating disagreements (Illinois State Board of Education [ISBOE], 2006).

The 2006 Technical Manual for MCA II reports training stage interrater correlations between .88 and .96 for the grade 8 mathematics extended response items. An adequate interrater reliability index collected at a training stage does not guarantee newly trained raters will rate stably over time during mass scoring. This is where the mass scoring process can be periodically monitored with SPC techniques. Also, with SPC charts raters can visually monitor their own level of rating quality on each batch of papers.

**Method**

In this study, we explored the use of SPC charts with a real classroom rating data set. Because SPC charts have not previously been used for mass scoring open-ended
items, we also provide a simulation study to explore the technique’s potential in this context.

Rules for Out-of-Control Processes

A mass production process is considered out of control or drifting out of control when at least one of the following (Montgomery, 1991, p. 117) is observed on any type of SPC chart (we illustrate the rules with a Shewhart $\bar{X}$ SPC chart in Figure 1):

1. Any point is outside the 3-sigma control limit (see batch 3 in Figure 1).
2. At least eight consecutive points are on one side of the chart (see batches 4 through 13).
3. Two or three consecutive points are outside the 2-sigma warning limit but inside the control limit (see batches 20 and 21).
4. Four or five consecutive points are beyond the 1-sigma limits (see batches 14 through 17).
5. An unusual or nonrandom pattern occurs in the data, such as a cyclic pattern (see batches 14 to 25 which form a quadratic pattern).
6. One or more points are nearing a warning or control limit (see batches 18 and 19).

The first condition listed above is the main rule for SPC charts. But to increase the sensitivity of these charts, with the passage of time the remaining rules have been added (Nelson, 1988). For an $s$ chart, however, violations at the lower warning or control limit do not carry the meaning of a severe process shift but rather a welcomed consistency.

Building a Shewhart $\bar{X}$ SPC Chart

For a Shewhart $\bar{X}$ SPC chart, past information or contractual specifications on the products’ mean and standard deviation are helpful to set up limits for the process. In industrial SPC applications, these means and standard deviations are often called “control” or “target” values. With these “control” values, the following steps are followed:
Figure 2. A Shewhart $\bar{X}$ control chart for rating mean differences on a statistics open-ended question.

1. The control mean $\mu$ can be used as the process center. This center signifies that the process is on target and that there is no disturbance such as drift or an abrupt shift.
2. A control standard deviation can be used to construct the process limits. The control standard deviation multiplied by 2 and divided by the square root of the batch size is used to obtain the upper and lower 2-sigma warning limits (i.e., $\mu_{X} \pm 2\sigma_{X}/\sqrt{n}$).
3. Similarly, obtain the upper and lower 3-sigma control limits (i.e., $\mu_{X} \pm 3\sigma_{X}/\sqrt{n}$).

The Shewhart $\bar{X}$ chart helps technicians monitor consistency in the process means. To use this Shewhart $\bar{X}$ chart for tracking mean differences, the same steps for building the $\bar{X}$ chart can be followed except that the target mean difference and the control standard deviation of rating difference are used in place of $\mu_{X}$ and $\sigma_{X}$, respectively.

Figure 2 provides a sample SPC chart where another rater provides ratings for open-ended papers scored by an instructor of an introductory business statistics course. The correlation between ratings is .909. The data size of 51 papers is too small to create "rational" subgroups through random sampling. Instead, we use the full data set with groups of size 3 papers as batches to illustrate the chart creation. For the rating difference, the instructor’s rating is subtracted from the second rater’s score. If the second rater is more stringent, the rating difference results in a negative score and vice versa. The goal of perfect rating agreement between rating sets is represented by a zero difference. Thus, a natural control difference of zero ($\mu_{d} = 0$) is used as the target for this SPC chart. In addition, because the open-ended question has not been used before, no past standard deviation can be used as the control value for the chart. So, a natural substitute is to use .594 which is the average standard deviation of the differences across batches. From Figure 2, there are two batches (11 and 16) nearing the warning limits and requiring further scrutiny.

Building an $s$ (Standard Deviation Estimator) SPC Chart

Often it may not suffice to monitor process means only because doing so does not guarantee that the individual products are also consistent. Thus, in addition to batch
mean stability, consistency in the within-batch variability is important and is often actively monitored in real industrial applications. Accordingly, alongside a mean chart, an $s$ chart is produced to see if product variability is fairly homogeneous between monitoring periods. To build the $s$ chart, the steps followed are:

1. The control standard deviation (from past information or contractual specifications) multiplied by a constant $c_4$ is used as the target, (i.e., $c_4 \sigma_X$);
2. The lower and upper warning limits are obtained by $(c_4 \pm 2\sqrt{1 - c_4^2})\sigma_X$;
3. The lower and upper control limits are given by $(c_4 \pm 3\sqrt{1 - c_4^2})\sigma_X$.

where $c_4 = \sqrt{2/(n - 1)}[\Gamma(n/2)/\Gamma[(n - 1)/2]]$, with the gamma function $\Gamma(n)$ defined as $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x} dx$ and recursion function $\Gamma(n + 1) = n\Gamma(n)$. Alternatively, $c_4$ can be simply approximated by $4(n - 1)/(4n - 3)$. The target and control limits are obtained from the fact that the expected value of $s$ is $c_4 \sigma_X$ with standard error equal to $\sigma_X\sqrt{1 - c_4^2}$ (Montgomery, 1991, p. 232). That is, $s$ is an unbiased estimate of $c_4\sigma_X$. To apply this $s$ chart for tracking the standard deviation of differences, the same chart building steps can be followed except the target standard deviation of rating difference is used in place of $\sigma_X$.

Figure 3 provides a sample $s$ control chart for monitoring the standard deviation of rating differences using the data from the introductory business statistics course. Again, no past standard deviation was available to be used as a control value for the chart. Instead, the average standard deviation of the difference score of .594 multiplied by $c_4$ of $4(3 - 1)/(4(3) - 3) = .8889$ was used as the control or target for the $s$ chart. Because the population standard deviation cannot be negative, if the lower limit is negative (as in this case) the lower limit is set to 0. Figure 3 shows 2 points outside the upper warning limit (batches 10 and 16) and 1 point (batch 6) approaching this limit. These batches can be highlighted for further scrutiny and follow-up.

Examination of individual papers, such as shown in Table 1, may be helpful to pinpoint the sources of discrepancies. Bold face values indicate mean and $SD$ differences flagged by the charts.
Table 1

<table>
<thead>
<tr>
<th>Batch</th>
<th>Instructor’s Ratings</th>
<th>Rating Difference</th>
<th>Mean Difference</th>
<th>SD Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>−1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>−1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>−1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>−1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. aRating difference = rater 2’s score—instructor’s ratings.

For example, batch 10 has an average mean difference of −.33, which is not flagged, whereas the standard deviation of rating differences of 1.1547 was flagged by the s chart for having larger than usual variation. Based on the Rating Difference column, this unusual variation specifically resulted from a slightly more lenient score on the first paper in the batch but a slightly more stringent score on the second and third papers by the rater, as compared to the scores by the course instructor. The results for the Shewhart mean chart are complemented by the standard deviation chart results and neither chart should be used in isolation from the other.

Mass Scoring Process: A Simulation Study

Unlike ratings for classroom tests such as in the previous example, mass scoring requires not only consistency in rating but also rater efficiency as there are many papers to score. It is important not to slow down or stop the scoring process unless absolutely necessary. However, because SPC charting techniques have not been applied to mass scoring in standardized testing programs, the article explores SPC application for mass scoring with simulation data.

The data in this article were simulated to follow technical characteristics of the typical grade 8 mathematics MCA II extended response item. Based on the interrater correlations cited above, a correlation of .92 was used as a parameter for the simulation. The MCA II follows a 5-point scoring rubric with 0 through 4 points assigned to any student paper. For this study, a set of scores by an expert rater was simulated. Another set of scores by a trained rater was simulated for the mass scoring of student papers.

Monitored papers. For illustrative purposes, 25 sets of 10 previously rated papers were chosen from past expert raters’ committee deliberations. It is assumed that as a natural outcome of this committee deliberation, the ratings of the expert raters are consistent across paper batches. These papers are termed “monitored papers” in this article. As such, from each batch, each of these 10 monitored papers already
had a score associated with it. In addition, these sets of papers contained student work from across the score range. The monitored papers were rated again by a rater at specific intervals in the mass scoring process. To clearly distinguish these monitored papers from other papers in the process, the other papers are termed “regular papers.”

**A mass rating scheme that is conducive to SPC application.** To periodically monitor the mass scoring process, the rating procedure followed the steps:

1. Select a batch of 10 previously expert-rated papers as monitored papers. The second rater rated these 10 papers and descriptive statistics were collected.
2. Subtract the initial or expert ratings from the second ratings rendered in the first step for these monitored papers. Graph the mean of the 10 rating differences.
3. Continue rating other batches of 10 regular papers and record their summary statistics. As usual, record the rating scores on individual papers.
4. When a specific number, say 90, of regular papers has been rated, rate another batch of 10 monitored papers. Collect descriptive statistics of the rating differences again and plot in the chart created in step 2. We chose the number 90 to coincide with the usual practice of choosing 10% of scored papers for a second reading (MDE & PEM, 2006; ISBOE, 2006).
5. Repeat the second through fourth steps and monitor the charted means of rating differences for any signs of nonrandomness or rating differences that were “out of control.”

The summary by batch of the rating differences on these monitored papers was captured and charted to detect any signs of rating drifts or aberrant ratings. If there was perfect agreement in ratings, the mean differences would be 0 with no within-batch variability. However, with sample idiosyncrasies within batch, one would tend to observe predominantly 0 mean differences with negligible within-batch variability. Substantially larger mean difference effects would be observed if a rater was either more or less lenient than the expert rating. In addition, substantially larger within-batch variability would be observed if large within-batch differences existed even if these differences, on average, cancelled out.

**Simulated data.** Data for the mass scoring process were simulated using MINITAB (2000) software. The performance rating data with different levels of disturbances were simulated for five raters over 25 monitoring periods. For each period, each rater rated a batch of monitored papers previously rated by an expert rater. The second ratings were simulated to correlate .92 with the expert ratings. Then the expert rater’s score was subtracted from the second rating. In this article, the population standard deviation of rating difference was assumed to be .60. This number was established by calculating the minimum standard deviation of the extended response items that would have resulted from the frequency distributions reported for the grade 8 MCA II mathematics assessments (MDE & PEM, 2006).

As discussed earlier, the interrater correlations reflect the quality of the mass scoring process. Because correlations by definition measure linear association between two variables, the data were simulated using the following linear regression model with details outlined later.
\( Y_r = \mu_r + \rho_{er}(X_e - \mu_e) + \phi\sqrt{1 - \rho_{er}^2}, \) (1)

where \( Y_r \) is a simulated (continuous) rating with mean \( \mu_r \) and variance \( \sigma_r^2 \), \( X_e \) is the simulated expert rating (also continuous) with mean \( \mu_e \), \( \rho_{er} \) is the correlation between the two ratings, and \( \phi \sim N(0, \sigma_e^2) \) is a random normal error term.

Some rating disagreements may follow nonlinear patterns and the correlation may be low due to nonlinearity in disagreements. Furthermore, for a 5-point scale a polynomial model up to degree 4 could be fitted. However, the main goal of the article is to provide an illustration of the techniques’ potential only. As such and to conserve space, we considered only the following quadratic regression model in the simulated data below.

\[ Y_r = \mu_r + \rho_{er}\left(\frac{X_e^2}{4} - \mu_e + .65\right) + \phi\sqrt{1 - \rho_{er}^2}. \] (2)

Because it contains a concave upward function with an expected value of re-ratings that is slightly lower than the expert ratings, the model generally represents rating stringency.

To illustrate the behavior of rating processes that were and were not in control, the following paragraphs provide a description of the parameters of the set of 10 ratings for each rater at each of the 25 monitoring periods.

Rater 1 was in control. The expert ratings for the monitored papers were sampled from a normal population with a mean of 2 and a standard deviation of .68 and were rounded to the 0–4 scoring scale. The mean and standard deviation were chosen to ensure representation of scores from across all the score ranges. The re-ratings for the same monitored papers were sampled from a normal population with a mean of 2 and a standard deviation of .68. To ensure an interrater correlation of .92 with the expert ratings, the linear regression relationship was used to derive the re-ratings. The re-ratings were also rounded to the 0–4 rating scale.

Rater 2 violated the upper control limit. Observed ratings were sampled like those for rater 1 except a very large disturbance was added to the rating difference at monitoring periods 18 and 19. The disturbance was normally distributed with a mean of .75 and standard deviation of .5.

Rater 3 violated the lower warning limit. Observed ratings were sampled as for rater 1 except a moderately large disturbance was added to the rating difference between monitoring periods 14 and 16 inclusive. The disturbance was normally distributed with a mean of −.35 and standard deviation of .25.

Rater 4 showed signs of nonrandomness. Observed ratings were sampled as for rater 1 except more negative runs of mean differences were sampled toward the end of the monitoring process. For periods 17–25 inclusive, instead of using the linear regression relationship, the stringent re-ratings were simulated using a quadratic regression model presented earlier.

Rater 5 showed signs of out-of-control variation with small-to-medium sized disturbances. Observed ratings were sampled as for rater 1 except between monitoring periods 2 and 5 inclusive and periods 22–25 inclusive, where there was a variability multiplication disturbance. A normally distributed disturbance with a mean of 0 and
standard deviation of the rating differences was added to the re-ratings to double the variability for these periods.

The expert rating on each paper was then subtracted from each rater’s scores. Then, the differences between the ratings were summarized by batch for further SPC analyses.

Building a Shewhart control chart for mean rating differences. Steps for building the Shewhart $\bar{X}$ chart explained earlier were followed. In particular, the mean rating difference of 0 was used as the center or target. In addition, the standard deviation of rating difference between an expert rater and another rater from a previous committee deliberation was used to obtain the limits of the chart. In this study, a standard deviation of .60 was used to mirror the minimum standard deviation of rating differences reported in the 2006 Minnesota MCA II Technical Manual (MDE & PEM, 2006). For example, the upper control limit in this article is $0 + 3(.60)/\sqrt{10} = .5692098$.

In addition to the consistency in mean differences between monitoring batches, consistency in the within-batch variability was important. Accordingly, for the monitored papers, an $s$ chart was produced to see if variabilities of rating differences were fairly homogeneous between monitoring periods.

Building an $s$ control chart for the within-batch variability of rating differences. To build the $s$ chart, steps for its construction explained earlier were followed. In particular, the standard deviation from the rater training of the rating difference between expert ratings and re-ratings of .60 was used to obtain the target and the limits of the $s$ control chart. Also because $n = 10$ in this article, $c_4 = \sqrt{2/(10−1)[\Gamma(5)/\Gamma(9/2)]} = \sqrt{2/9(24/11.632)} = .97266$. Alternatively, $c_4$ could be approximated as $4(10−1)/(40−3) = .97297$.

Results

Rating differences of various degrees of quality were analyzed and reported using the SPC techniques discussed in this article. The analyses are summarized in Figures 4 through 8 and Table 2.

Figure 4a provides a Shewhart control chart of the mean rating differences by rater 1 versus the expert ratings over several quality control periods. Figure 4b provides the corresponding standard deviations of rating differences. As explained in the previous sections, it is not difficult to see in Figure 4a that because no points are beyond the warning limits and the pattern is random, the mean rating differences for the monitored papers are on target and are stable over time. From a statistical inference perspective, this also indicates some comfort level regarding the regular papers rated between these quality-controlled batches because drifts, if present in a process, will tend to persist until corrective actions are taken.

Figure 4b shows an $s$ chart for the standard deviation of rating differences by rater 1. On the surface, it might seem as if there are many violations at the lower control limit. However, these represent 0 standard deviations that imply that ratings by rater 1 are consistently identical to those by the expert rater. That is, for performance assessment applications, this is not alarming news; rather, a welcomed one.
For the $s$ chart in the current application, we are concerned if violations occur at the upper warning or control limits as these suggest highly inconsistent within-batch ratings across batches. There seem to be no violations in Figure 4b so ratings by rater 1 seem consistent and in statistical control.

Figures 5–7 provide similar Shewhart mean control charts but by raters 2, 3, and 4, respectively. Figure 5 shows signs of flagrant violation of the Shewhart mean chart for monitoring periods 18 and 19, as circled on the graph, because these mean differences are above the 3-sigma upper control limit. Figure 6 shows violations of the 2-sigma lower warning limit for periods 15 and 16 as circled on the chart. In addition, batch mean difference for monitoring period 14 shows the initial alarming sign of approaching the 2-sigma lower warning limit. Finally, Figure 7 shows signs of nonrandomness from periods 18 to 25 because the mean rating differences, as circled on the chart, are consistently on the lower side. This indicates that, on average, rater 4’s scoring is drifting and is relatively more stringent than those by the expert rater. To conserve space, $s$ charts for raters 2 through 4 are not provided because their rating difference standard deviations look similar to those for rater 1.

Figure 8b shows an $s$ chart of rating differences for rater 5. Periods 2, 4, and 23 to 25 show larger variation than expected as these standard deviations are either outside the 3-sigma upper control limit or outside the 2-sigma upper warning limit of an $s$ chart. In addition, periods 3, 5, and 22 show standard deviations approaching...
the 2-sigma upper warning limit. These large “out-of-control” standard deviations indicate that rater 5 rated differently from the expert rater at a very inconsistent rate. That is, for these batches rater 5’s scores are not consistently, but variably, lower or higher than the expert rater’s ratings. For batch 2, the standard deviation is large whereas the Shewhart mean chart in Figure 8a shows zero mean difference. This signals that large negative rating differences totally cancel positive ones. Although they have mean differences within the limits of the Shewhart mean chart, batches 4, 23, 24, and 25 violate warning or control limits of the s chart. This indicates that, for these batches, large rating differences exist and are partially masked by cancellation. In addition, batch 22 has a mean difference that violates the warning limit of the Shewhart SPC chart and standard deviation nearing the warning limits of the s chart. For these batches, the typical course of action is to check the individually rated papers. Associated papers within batches can be checked (such as with Table 1 earlier) to see which paper causes the most rating inconsistency.
Thus far, our discussions have centered on the SPC charts. However, for educational measurement settings, the discussion appears inadequate without also examining the interrater correlations in relation to these SPC results. For this, the interrater correlations were obtained when papers, and paper batches, were the analysis units. This was done to aid the discussion of the between- and within-temporal consistency of rating differences.

Table 2 shows the corresponding interrater correlations for the simulated ratings when the units of analysis are individual papers and batches of papers. In addition, the interrater correlations are calculated when there are disturbances and no disturbances of the type and magnitude explained earlier. The correlations are all significantly different from 0 at the .01 level with \( p \)-values less than .001.

The first two columns of the table provide these correlations when the unit of analysis is the individual paper. For rater 2, when some of the rating batches violate the 3-sigma control limit of the Shewhart chart for the mean rating differences, the interrater correlation is depressed to .871 compared to .923 when there were no violations. Similar but milder effects can be seen with rater 3 where some batches violate the 2-sigma warning limits of the Shewhart mean chart. Although ratings by rater 4 do not violate the 2-sigma warning limits of the Shewhart mean chart, consistent drift on the stringent side produced similar effects as those shown by rater 3. Ratings from rater 5 appear to show the most severe effects on the correlation where variability disturbances are present. When the 2-sigma limits are violated and when nonrandom patterns (drift) are observed, the interrater correlations are still in the .90s. However, when the 3-sigma limits are violated and when the standard deviations are significantly large, the correlations fall below the usual .90 acceptability level.

The last two columns of Table 2 provide these same correlations but when the unit of analysis is the paper batches. Generally, when there are no disturbances, except for rater 3, the correlations are higher than when individual papers are the units of analysis. When there are some disturbances, except for rater 5, the correlations are depressed further (below .90). One thing to note is that, as shown...
Table 2

*Interrater Correlations across Individual Papers and Batches of Papers with Varying Disturbance Levels*

<table>
<thead>
<tr>
<th>Rater</th>
<th>Across Papers</th>
<th>Across Paper Batches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Disturbance</td>
<td>Some Disturbance</td>
</tr>
<tr>
<td>1</td>
<td>.939</td>
<td>.956</td>
</tr>
<tr>
<td>2</td>
<td>.923</td>
<td>.871</td>
</tr>
<tr>
<td>3</td>
<td>.928</td>
<td>.903</td>
</tr>
<tr>
<td>4</td>
<td>.925</td>
<td>.896</td>
</tr>
<tr>
<td>5</td>
<td>.922</td>
<td>.716</td>
</tr>
</tbody>
</table>

On the SPC charts, the disturbance for rater 5 is of the variability type whereas for the others the disturbances are in the mean rating differences. As groups (or batches) are typically less variable than individual papers, this tends to explain why we observe the results provided in Table 2. For the correlation analysis reported in the last
two columns, because the means of the batches are the only input, within-batch variability is removed. That is, when grouping by paper batches, variability disturbances within batches (such as in rater 5’s ratings) are removed, whereas disturbances in the mean ratings across batches are more pronounced as the violating batches gain more prominence in the batch analysis.

Table 2 confirms Brennan’s (1995) findings on the reliability of group means where school means are not necessarily more reliable than student scores. In this study batch means were not more reliable than individual scores when the main source of temporal inconsistency, as indicated by Shewhart mean charts, was between batches but were more reliable when the main source of temporal variability, as indicated by the s chart, was within batches.

**Discussion**

Clearly, it is possible to display two types of temporal consistency information through SPC charts that each rater can be made aware of by communicating the level of rating quality. Thus, corrective actions can be taken to improve rating quality from time to time. Also important is that, with current computer technology, such informative charts can be produced simultaneously while raters score papers. Without such charts to guide them, raters may continue rating while being oblivious to the hazards of drifting. It is also possible to temporarily stop mass scoring for corrective actions when some indications of “out-of-control” ratings are observed. Thus, SPC charts such as those we have demonstrated in this article seem to hold promise for monitoring stability of performance ratings over time.

Correlations and percentages of agreement, which typically have been the only indicators of interrater reliability, globally measure the extent of agreement between ratings for the whole mass scoring process. They do not show where or when specific rating disagreements occur. In particular, when disagreements between ratings are consistent, the correlation can be misleadingly high. In contrast, the SPC charts signal where rating discrepancies occur during the rating process rather than after. This provides an advantage in that while ratings are being conducted the rating process is monitored as a live feedback for raters to help improve process consistency. The main aims in using the SPC charts are to increase the consistency wherever possible and to decrease inconsistencies while they happen without the unnecessary cost of slowing or stopping the mass rating process.

As indicated in the introduction section, SPC techniques have been extensively used in mass production in business and industry. As the main driving force of the current International Organization for Standardization (ISO) series of standards (e.g., Hoyle, 2001), these techniques have been very useful in the rapid development in the industrial sector of the economy. It is hoped that with this article, readers can also see the usefulness of such SPC techniques in the field of educational measurement.

**Limitations and Further Research Directions**

Although quality control of randomly assigned monitored papers can be accomplished as outlined in this article, most papers are not rated a second time. Thus, like mass production in industrial settings, there is a possibility that some aberrant
ratings may creep into the mass scoring. But with initiatives for monitoring scored papers, the possibility of “out-of-control” rating drifts may be minimized as batches of regular papers immediately adjacent to an “out-of-control” monitored batch could be randomly checked as well.

Other statistical graphs such as histograms of rating differences across individual papers can also be used to provide some idea of the distribution of discrepancies and agreements. An advantage of these methods is that they require checking only one graph. However, as these graphs are for individual papers, unlike SPC charts discussed in this article, they have no standard errors to establish warning systems for early detection of large rating discrepancies and nonrandom patterns (or rating drift).

To get around this issue, a box plot on mean rating differences with standard error bars can be constructed. One disadvantage of this approach is that the targets and control limits for the standard deviation of rating differences cannot be plotted on the same chart. Hence, using this method, standard deviations of rating differences cannot be monitored simultaneously. Another disadvantage is that the reason for flagging large rating differences will not be clearly due to either large mean differences only or simultaneously large mean and large standard deviation differences. In contrast, although two control charts need to be simultaneously checked, the reasons for flagging large rating differences will be clear because each chart will have its own standard error and control limits. In addition, unlike other techniques and only when an out-of-control process is observed, momentary stopping of the mass scoring process for corrective actions may improve both consistency and rating efficiency. Moreover, some means flagged by the Shewhart chart (e.g., batch 11 in Table 1) also include a recurring pattern of 1-point discrepancies. This aberrant pattern is currently accepted in many state programs if other non-SPC methods such as the histogram method are used.

In the method section of this article, the recommendation of 10 monitored papers in step 1 of the SPC chart construction and 90 regular papers between successive monitoring periods in step 4 are partially based on current scoring industry practices. However, the rating scheme discussed in this article has not been test driven yet with live operational data to provide insights into “what works best” in practice. That is, because this is one of the first studies on implementing SPC techniques in the area of performance assessments, no statistical backing for the numbers of 90 regular and 10 monitored papers has yet been produced. For example, which 10% monitoring scheme is most reasonable for a single rater to handle? Is 25 batches of 100 papers (i.e., 2,500 papers total), 25 batches of 30 papers (27 regular plus 3 monitored), or some other scheme more reasonable? Because this question requires an additional empirical study, preferably in an operational setting, we urge other researchers to pursue the best empirical number of regular papers between successive re-ratings of monitored papers while considering economic and pragmatic limitations.

There are some other related questions that cannot be answered in a single introductory study such as this and may be worth pursuing. For example, the effects of changing raters on the monitored batch statistics of the Shewhart mean and \( \bar{s} \) charts are still not known. This question is particularly important if these raters are not really interchangeable. Accordingly, the question on the practical
maximum and minimum number of monitored batches for a single rater can also be
investigated.

Also, simulated data can be generated alternatively through other nonlinear mod-
els to represent some possible rating disagreements that may occur in practice. How-
ever, because the main intent of this article is only to provide an illustration of the
SPC techniques’ potential, to economically achieve this we used regression to sim-
ulate the data. If other models were used, some of the results reported in this study
may have been somewhat different.

In this article, the design used is rater by papers within batches. Omar (2007) has
discussed the technique for the design where the same rater rates the same papers
within batches on two separate occasions. Nevertheless, application of these SPC
techniques to other data collection designs still needs to be researched.

Application of SPC techniques to a live mass-scoring process such as the MCA
II can further enrich the literature. However, because SPC charts have not yet been
applied to such processes, the techniques’ merits and drawbacks to empirical data
are not fully known. Without this knowledge, stakeholders of these performance as-
se ssments may not be as receptive to trying these SPC techniques. Thus, simulation
work that can shed light into the merits and drawbacks of these techniques is needed
to convince stakeholders to use them. Hopefully, this article is an initial step in this
direction.

Acknowledgments

The author would like to express gratitude to Mr. A.R. Congreve for providing pro-
fessional editing help on language mechanics for the article and to three anonymous
reviewers and the JEDM editor for insightful comments to strengthen the article. The
author also would like to acknowledge the research facilities at KFUPM.

Note

1A short proof of these statistical facts for the s distribution can be made available from the
author upon request.

References

American Society for Quality Control—Automotive Division & Automotive Industry Action
Southfield, MI: Authors.
cational Measurement, 32, 385–396.
Burr, I. W. (1967). The effect of non-normality on constants for $\bar{X}$ and R charts. Industrial
Quality Control, 23, 563–568.
the American Mathematical Society, 28, 427–432.
Deming, W. E. (1951). Elementary principles of the statistical control of quality—series of
lectures. Tokyo, Japan: Nippon Kagaku Gijutsu Remmei.
SPC Charts for Temporal Consistency


Author

M. Hafidz Omar is Assistant Professor of Statistics, Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; omarmh@kfupm.edu.sa. His primary research interests include statistical process control, IRT vertical scaling, distribution theory, and statistical methods in psychometrics.