Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 15, October 2011 ISSN 1531-7714

Quality Control Charts in Large-Scale Assessment Programs

William D. Schafer, *University of Maryland*

Bradley J. Coverdale, *George Washington University*

Harlan Luxenberg, *University of Maryland*

Ying Jin, *United BioSource Corporation*

There are relatively few examples of quantitative approaches to quality control in educational assessment and accountability contexts. Among the several techniques that are used in other fields, Shewart charts have been found in a few instances to be applicable in educational settings. This paper describes Shewart charts and gives examples of how they have been used to monitor quality in testing programs. Additional areas of application in large-scale educational assessment programs are proposed

Quality Control Charts (QCC), also called Statistical Quality Control and Acceptance Sampling, have historically been used to monitor product quality in a production or manufacturing environment. Their general purpose is to provide information that can be used to uncover discrepancies or systematic patterns by comparing expected variance verses observed variance. In a production environment, that propose translates to improving product quality and productivity in order to maximize a company's profits. Deming, who was a major contributor to quality control research and whose techniques have been credited with reviving the Japanese automobile industry, believed that the quality of a process can be improved using QCCs (Deming, 1982).

Technicians visually inspect QCCs to determine if deviations from an expectation fall outside certain bounds, if there are any systematic patterns that appear on the chart or if the points fall very far from the expectation. If any of these situations are observed, then the process is considered "out of control." Some variability is normal and can be caused by sampling fluctuations and by differences among sampled groups. If the fluctuations appear within established outer bounds and the pattern of deviations appears to be random, then the process is considered "in control." When this happens, no investigation is conducted on the data since the observed process variations are expected.

There are many different variations of control charts that can be used to detect when processes go out of control. These are described in detail in Basseville and Nikiforov (1993). The most common and easily interpretable of these is the Shewart control chart. These charts, named after Walter Shewart, were created from an assumption that every process has variation that can be understood and statistically monitored (Savić, 2006). A Shewart chart includes three horizontal lines, a center line, an upper limit, and a lower limit. The center line serves as a baseline and is typically the expected value or the mean value, while the upper and lower limits are depicted by dashed lines and are evenly spaced below and above the baseline.

A control chart is essentially a graphical interpretation of a standard, non-directional hypothesis test. The hypothesis test compares each point on the chart with an in-control range. If a point in the control chart falls within the upper and lower bounds, it is akin to failing to reject the null hypothesis that the process is in-control. A point that falls outside the bounds can be thought of as the same as rejecting the null hypothesis. Type I and Type II errors also have analogs in using a control chart. Determining that a process is out of control when it is really not is analogous to a Type I error and accepting that a process is in control when it really is not is analogous to a Type II error. We note in passing that an equivalency interval based on either empirical history or logic might be established for the statistic and significance testing proceed as suggested by Rusticus and Lovato (2011).

A control chart utilizes a measure of central tendency for the baseline and a measure of variability for the control limits. The most common of these charts are the M, S, and

R charts. These refer to mean, standard deviation, and range respectively. Time (or occasion) of the sample can be plotted on the horizontal axis of the chart and the observation taken from the sample on the vertical axis. For each chart, the following three things must be decided before they can be created: how often the samples will be drawn, how large the sample will be, and what will be used as the control line and the control limits.

In order to use a QCC, a sample is drawn from a population of scores and then some characteristic of it is plotted. In the production environment, this might mean selecting a small sample of produced units every hour. A visual inspection of these graphs allows an engineer to quickly inspect the quality of the current production run. Thought must be given to both how often a sample should be selected and the size of the sample. In general, the larger the sample, the better chance changes or variations in the process will be noticeable (Montgomery, 1985). The most beneficial situation would be to have a large sample frequently selected for measuring in the control charts. This is often not very feasible due to data and economic restraints, so some combination of sample size and frequency that the sample is drawn must be selected for each study.

In order to set the upper and lower bounds, a common procedure is to set them three standard deviations (sigmas) away from the baseline (although they can be set to different values based on the process). The value of σ_x can be determined using the previous observations on X. In order to set the limits three sigma away, the below formulas would be used:

Upper Control Limit = $\mu_x + 3\sigma_x$

Baseline = μ_x

Lower Control Limit = $\mu_x - 3\sigma_x$

A QCC may be thought of as an accumulating history of the statistic being charted. Its purpose is to detect when something unusual has happened, and that is taken as an indication that the process being described may not be behaving normally, or may be "out of control." An operational definition of "unusual" is needed. Although various operational definitions exist, there are six main criteria that are commonly checked each time a new data point is added. If any one (or more than one) of these criteria is met, then the process may be out of control (Montgomery, 1985).

1. The most recent point is outside of the upper and lower control limits.
2. At least seven consecutive points are on the same side of the baseline.
3. Two out of three consecutive points are outside a 2-sigma line.
4. Four out of five consecutive points are outside of a 1-sigma line.
5. Any pattern that is noticeable that is non-random or is systematic in any way.
6. Several points are close to the upper or lower control limits.

QCC charts, while common in business, have only recently been used in education, Omar (2010) cites a few educational studies using control charts for determining IRT parameter drifts in a computer adaptive environment as well as developing a person-fit index. But it is rare to find them used for monitoring statistical characteristics of state or other achievement testing programs. In this paper, we describe some uses that have been made of Shewart charts in a practical environment and suggest some further uses.

Methodology

In this section we describe how Shewart charts have been used in a large-scale assessment program. Several years ago, Maryland tested all students in grades 3, 5, and 8 using a performance assessment format in reading, writing, math, science, language usage, and social studies. The program was unique in design and required both innovative assessment development and novel applications of existing designs. Among the novel applications was the use of equating methodology to link each year's test forms to the original scale scores. Each step of the process was carried out by the state's vendor, who recommended to the state whether to proceed to the next step. The state was advised by a group of nationally recognized psychometricians to help reach a decision, and in later years the group was informed when the results were judged unusual, using QCCs.

Maryland's National Psychometric Council (NPC) began to use QCCs in 2001 to help determine whether or not to recommend accepting the scaling and linking work of its contractor for the Maryland School Performance Assessment Program (MSPAP). The state contracted at that time with the Maryland Assessment Research Center for Education Success (MARCES) at the University of Maryland, College Park to create QCCs based on several years of contractor reports and to report those that were out-of-range to the NPC.

The MSPAP, consisting entirely of constructed-response items, was administered in three forms, referred to here as operational clusters A, B, and C. Clusters were randomly distributed within schools across the state. Each cluster measured all six content areas: reading, writing, language usage, math, science, and social studies. However,

the clusters were not parallel; although all of the content areas were assessed across all three clusters, the clusters did not sample the content equivalently. The results of the clusters taken together were used to assess school performance. In selected schools, a fourth cluster, called the equating cluster, was also included in the randomization; this cluster was repeated from one of the previous year's clusters. The papers from a sample of students who took the equating cluster the prior year were scored in the current year to provide data to adjust for rater differences between the two years. After initial independent calibrations of the three operational clusters, linking using the linear equi-percentile technique and the two-parameter, partial-credit model (2PPC) proceeded in three steps: (1) one of the three operational clusters was chosen as a target and the other two were linked to it; (2) the target cluster was linked to the equating cluster; and (3) the equating cluster results from the current year were linked to those from the prior year. Results from the three steps (the first also included the independent calibrations) were reported to the state and discussed by the NPC separately.

MARCES computed and developed QCCs each year for these quality indicators for MSPAP. Those that were out-of-range were reported to the NPC. The budget for this work was under $20,000.

Descriptions of statistics for control charts

The following statistics were based on the calibration output (step 1: initial calibration and linking for the three operational clusters), year-to-year linking output (step 2: linking the main cluster to the equating cluster), and rater-adjustment output (step 3: linking the current scorers with the prior scorers). MARCES developed control charts for these statistics based on the year-to-year data for quality control purpose. Each is either an original statistic present on the output or was computed.

All these were calculated on each grade-level and content area combination. There were three grade levels and six content areas, for a total of 18 combinations. Only those control charts that were "out of range" were forwarded to the NPC, allowing them to focus only on the results that were unusual in any one year.

a. Based on the calibration output for each operational cluster separately:

1. Alpha – reliability coefficient
2. Mean and standard deviation of item discrimination parameters
3. Proportion of 2PPC category threshold reversals: number of item threshold (g) patterns other than $g_1<g_2<g_3<g_4$ … divided by the total number of items (where g_1 is the equi-probability point for scores of 0 and 1, etc.).
4. Mean and standard deviation of an item-fit statistic
5. Off-diagonal r: average inter-correlation of the item residuals, controlling for person ability
6. Proportion of r>0: number of positive inter-correlation coefficients of the item residuals divided by the total number of item residual correlations

b. Based on the test cluster linking output:

1. Difference between highest and lowest means – the difference between the highest and lowest means of item-pattern scores among the three clusters.
2. Difference between highest and lowest sigmas– the difference between the highest and lowest sigmas of item-pattern scores among the three clusters.
3. The largest conditional standard error of measurement (CSEM) at the Lowest Obtainable Scaled Score (LOSS) and Highest Obtainable Scaled Score (HOSS) among the three clusters.
4. The largest percentage of students (IP%) at the LOSS and HOSS based on item-pattern scoring among the three clusters.
5. Difference between largest and smallest percentiles across the three clusters at the cut score between proficiency levels 2 and 3 (level 2 can be thought of as "advanced").
6. Difference between largest and smallest percentiles across the three clusters at the cut score between proficiency levels 3 and 4 (level 3 can be thought of as "proficient").
7. Proportion of scores at the LOSS and HOSS – number of students at the LOSS and HOSS divided by the total number of cases, for each cluster.
8. The largest CSEM at the 2 vs. 3 and 3 vs. 4 cut scores.

c. Based on the equating cluster linking output:

1. The difference between the means of item-pattern scores between the two clusters (target vs. equating).
2. The difference between the standard deviations of the item-pattern scores between the two clusters (target vs. equating).
3. The effect size between the two clusters; the effect size 'd' is computed as follows:
   $d = (mean_1\text{-}mean_2)/S_p$
   S_p = pooled standard deviation
4. The larger CSEM at the LOSS and HOSS.

5. The larger percentage of students at the LOSS and HOSS based on item-pattern scoring (IP%) between the two clusters.
6. Difference of the percentiles at the proficiency level of 2 vs. 3 cut score.
7. Difference of the percentiles at the proficiency level of 3 vs. 4 cut score.
8. Proportion of scores at the LOSS and HOSS – number of students at the LOSS and HOSS divided by the total number of cases for each of the two clusters, for each cluster.
9. The larger CSEM at the proficiency level of 2 vs. 3 and 3 vs. 4 cut score.

d. Based on rater-linking output:

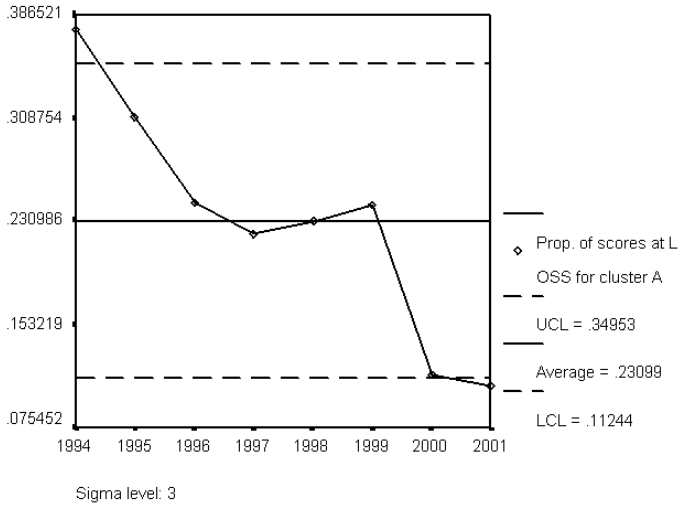
1. The difference between the means of item-pattern scores between the two clusters.
2. The difference between the sigmas of item-pattern scores between the two clusters (target - equating).
3. The difference of the raw score means between the two rater groups.
4. The difference of the raw score sigmas between the two rater groups.
5. Standardized raw score mean differences (effect size) between the two rater groups.
6. The larger CSEM at the LOSS and HOSS between the two rater groups.
7. The larger percentage of students at the LOSS and HOSS based on item-pattern scoring between the two clusters.
8. Difference of the percentiles at the proficiency level of 2 vs. 3 cut score between the two clusters.
9. Difference of the percentiles at the proficiency level of 3 vs. 4 cut score between the two clusters.
10. The larger CSEM at the proficiency levels of 2 vs. 3 and 3 vs. 4 cut scores.
11. The number of students at the LOSS and HOSS divided by the total number of cases for each of the two clusters

Below are five examples of QCCs for variables that were reported out of range in 2001. (All were in-range in 2000, which can be taken as examples of in-control charts.) The prior data were taken from the identical statistics computed on prior years, beginning with 1996, the first available. While the NPC recognized that these variables could have scores outside of the six-sigma range (the most common range) due to chance, any patterns found were used for further discussion about the equating of the assessment. All figures were developed using SPSS at its default values (upper and lower limits three standard deviations from the mean).

1. Proportion of Scores at Lowest Obtainable Scaled Score (LOSS): Writing Grade 3 Test, Cluster A

In this example, there was an unusually low proportion of scores at the lowest obtainable scale score (LOSS). The historical range was between approximately 11%-33%, with an average of about 23%. In 2001, only 11% of the writing scores were at the LOSS. The NPC concluded that this was indeed a desirable trend since students scoring at the LOSS could only occur through poor achievement or poor measurement and a decrease in the proportion at the floor may be taken as a result of education success. Since the prior year was also low, this trend is desirable and a new baseline and lower control limit (as well as upper) may result in the future.

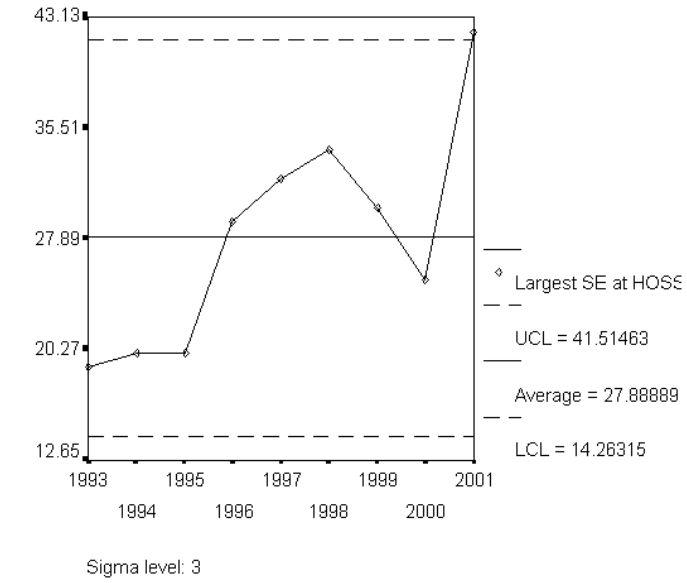
Figure 1: Proportion of scores at LOSS for cluster A over time



2. Standard Error of Highest Obtainable Scaled Score: Language Usage

The contractor included in their work the conditional standard errors (CSEMs) of all scale scores for each cluster. MARCES generated QCCs for the largest CSEM for several points, including the LOSS and the HOSS. In this case, the CSEM for the HOSS fell outside the range for Language Usage. The statistic was slightly higher than the typical range. In this case, the NPC did not recommend any action since it seemed like an isolated and mild example.

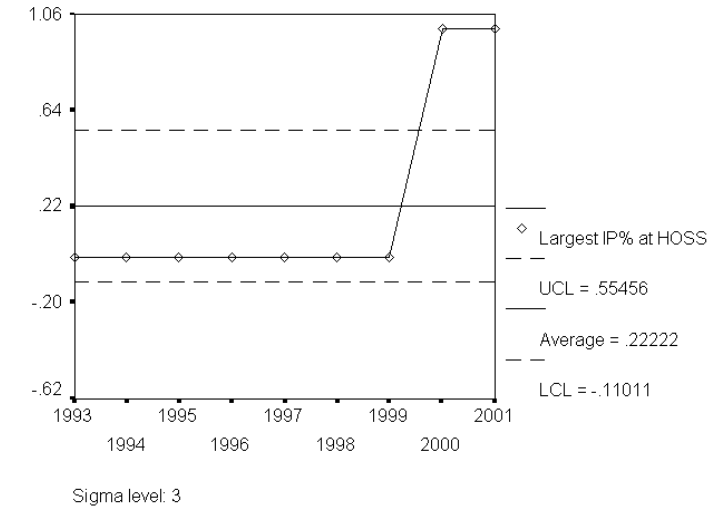
Figure 2: Largest Standard Error at HOSS over time



3. IP% at the HOSS: Science, Grade 8

Using item pattern (IP), or maximum likelihood scoring, the largest proportion of students at the HOSS among the three clusters was tracked. This control chart shows a remarkable pattern in that a stable percent over the first few years changed to what appears to be a new stable pattern in the last two years. Although this pattern is a positive indicator for the state (more students scoring in the upper ranges), the stability raises concern that it is an artifact. The NPC recommended watching this statistic in the future to see whether further investigation may be needed.

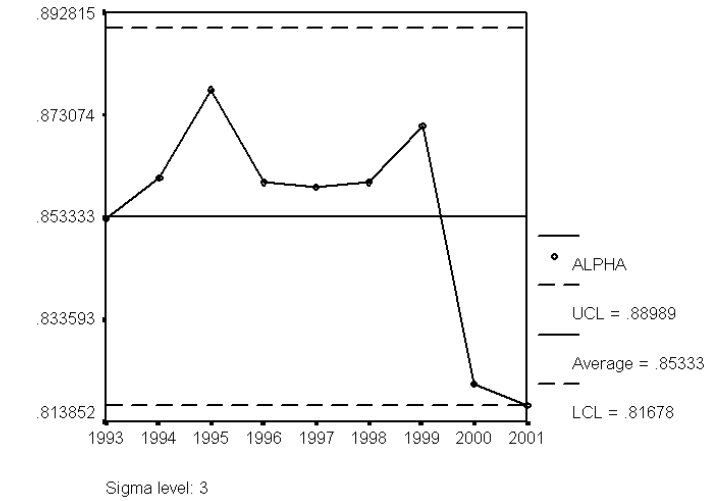
Figure 3: Largest proportion of students at HOSS over time



4. Reliability of Cluster A: Reading, Grade 8.

One of the measures of reliability reported by the contractor was the alpha for each cluster for each content area. The alpha for Cluster A in 2001 was below the lower control limit. The NPC noted that the alpha for Cluster B was even lower but not out-of-range. The flag in the example seems to be the result of a series of high and consistent alphas for cluster A's in 1993-1999. But the combination of this out-of-range result and the even-lower alpha for cluster B does seem unusual. The overall reliability for Grade 8 reading may be unusually low in 2001.

Figure 4: Reliability over time

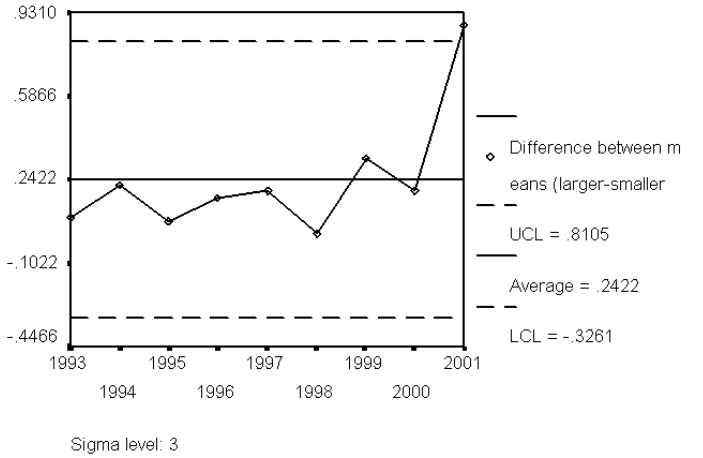


It should be noted that fairness implies that alphas should be neither too low nor too high across assessment forms. Especially for a high stakes examination such as admissions or certification (which this is not), it is not fair for those who are assessed on a form with more measurement error to be compared with those who are assessed on a form with less.

5. Difference in item-patterns scores: Math, Grade 3

QCCs were used to monitor the difference in the means of the item-pattern scores across clusters. Since the three clusters were randomly distributed within schools, the differences should reflect only chance variation. In this case, the means seemed more varied for grade 3 math in 2001 than they were for earlier years. The NPC did not find any anomaly that would explain this observation.

Figure 5: Difference between means over time



DISCUSSION

Our work suggests that the use of QCCs can be used to monitor the quality of an educational assessment program and to make efficient use of technical advisors' time by focusing their attention on the more unusual findings. But QCCs can monitor other qualities than contractor analyses in large scale testing environments as well. While experts in the field will determine acceptable baselines and control limits, we mention a handful of topics and examples in which control charts might be beneficial for monitoring quality in large-scale testing environments.

A. Proportion of field test items that make it to approval for use as operational items.

It is important for a test bank to have many similar questions that measure the same learning objective. Not only does this help in terms of comparing similar forms, but this also helps to ensure that students are able to demonstrate the knowledge breadth and depth required in the content standards. Having many questions with similar characteristics also helps to ensure consistency among forms. Thus, it could be useful for a testing program to monitor the proportion of test questions that make it to approval. If the proportion accepted falls below the observed range, then the process may be becoming inefficient and further study may suggest corrective actions.

B. Item Analysis of Scores

In addition to investigating the number of questions created and used each year, one can also evaluate various elements of an item analysis using QCCs. The item difficulty (p-value or IRT b-value or other location measure) and the item discrimination (correlation or IRT a-value) could be useful results. Both the mean and standard deviation of these statistics could be plotted for the field tested items. Changes in these statistics over time may have implications for changes in item development activities.

C. Proportions Above Cut Scores

With many policy decisions occurring because of a student's performance compared with cut scores, it is imperative to determine whether proportions of students in the various achievement levels are consistent with past trends. Using a control chart for each cut score, the changes in overall population (or subgroup) results could be compared with past outcomes to study whether trends have been broken, at state, district, or school levels.

D. Linking Block Characteristics

Assuming there is a candidate linking pool of items that have been given for that purpose and assuming they are evaluated to cull those that are acceptable as linking items, the proportion rejected could be charted. In any event, the correlation between the linking items "subtest" and the rest of the operational items as another "subtest" could monitor the quality of both items sets over time.

E. Item Block Positioning

Some argue that the location of a test item will affect a student's answer, whether due to time management issues while testing, fatigue, or other factors. In order to chart location, one could find the total number of items that were used on at least two forms, as well as what timing blocks that they were located in, and find the difference. The averages and standard deviations of the differences could then be plotted across administrations of different forms in the same year or across years.

F. Statistical Characteristics of Test and Subtest Scores

The reliabilities and inter-correlations of tests and subtests could be helpful characteristics to monitor. An interesting possibility might be the conditional standard errors of measurement on each of these at the various cut scores.

G. Item Awakening

Another aspect to investigate might be how long an item is not used before it is "awakened." Items that are awakened each testing period would need to be identified along with the time that had passed while the item was not in use. Mean and standard deviation of the time frames for each subject could be plotted in separate charts across each test administrations.

Conclusion

By investigating each of these criteria using control charts, testing programs may be able to spot trends that have

occurred across time and make decisions about whether further investigation is needed when out-of-range observations are found. The examples used in this paper demonstrate that QCCs can be useful even for relatively new programs. It is suggested that the technical advisory group for an assessment program be presented with the charts on a routine basis so they can decide if the data requires cause for concern and what recommendations they have for addressing any raised issues.

While we understand that QCCs can be applied to education via large-scale assessments, there is a gap in the research regarding the minimum data requirements that are needed to ensure the QCCs can be useful in practice. As testing programs decide to report on their experience and findings from using QCCs, new parameters can be created that apply better to the education sector, instead of simply relying on business precepts. We anticipate these and other questions can be addressed as QCCs receive more attention in assessment programs.

REFERENCES

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes - Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall.

Deming, W E. (1982). *Out of the crisis*. Cambridge, MA: MIT Press.

Montgomery, D. (1985). *Introduction to statistical quality control*. John Wiley & Sons: New York.

Omar, H. M. (2010). Statistical process control charts for measuring and monitoring temporal consistency ratings. *Journal of Educational Measurement*, *47*(1), 18-35.

Rusticus, S. A. & Lovato, C. Y. (2011). Applying Tests of Equivalence for Multiple Group Comparisons: Demonstration of the Confidence Interval Approach. *Practical Assessment, Research & Evaluation*, 16(7). Downloaded August 8, 2011 from http://pareonline.net/getvn.asp?v=16&n=7.

Savic, M., (2006). P-charts in the quality control of the grading process in the high education, No 200636, Working Papers, Faculty of Economics in Subotica

Citation:

Schafer, William D., Coverdale, Bradley J., Luxenberg, Harlan & Jin, Ying (2011). Quality Control Charts in Large-Scale Assessment Programs. *Practical Assessment, Research & Evaluation*, 16(15). Available online: http://pareonline.net/getvn.asp?v=16&n=15

Acknowledgement

This paper was partially funded by the Maryland State Department of Education (MSDE) through the Maryland Assessment Research Center for Education Success (MARCES) at the University of Maryland. The opinions expressed are those of the authors and not necessarily those of MARCES or MSDE. The authors are indebted to MSDE for sharing the statewide assessment results necessary to complete the work described here.

Authors:

William D. Schafer is Affiliated Professor (Emeritus), Maryland Assessment Research Center for Education Success, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD. He specializes in assessment and accountability.

Bradley J. Coverdale is a higher education administration doctoral student at George Washington University. His research interests include investigating the effects of proprietary education, and access and persistence issues for first generation college and minority students.

Harlan Luxenberg is a graduate assistant in the Maryland Assessment Research Center for Education Success, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD.

Ying Jin is Research Scientist, Outcomes Research, United BioSource Corporation (UBC), Bethesda, MD. Prior to joining UBC, she was Senior Research Analyst at the American Institutes for Research, Washington, DC. She specializes in large-scale assessment and survey methodology.