THE USE OF STATISTICAL QUALITY CONTROL CHARTS TO EVALUATE CHANGES IN INDIVIDUAL PERFORMANCE

Randa L. Shehab and Robert E. Schlegel
School of Industrial Engineering

Kirby Gilliland
Department of Psychology

University of Oklahoma
Norman, OK 73019

An employee's readiness to perform (RTP) has become an important issue facing today's industries. Some industries have turned to cognitive performance testing to provide answers regarding their employee's abilities to work safely and effectively. Such tests are designed to assess the employee's current state of preparedness for work without identifying specific causes for any noted performance impairment. This paper evaluates performance-based RTP tests with regard to the metrics by which performance change is judged. In addition to evaluating a current commonly used metric, several statistical quality control charts were examined as alternate methods for identifying impaired performance. Traditional Shewhart charts were used as well as exponentially weighted moving average charts and cumulative sum charts. A comparative analysis of the various methods revealed that control chart techniques provided superior effectiveness over the current method. Specifically, exponentially weighted moving average charts were effective in evaluating continuous performance measures and Shewhart p charts were effective in evaluating discrete measure data.

INTRODUCTION

Readiness-to-perform (RTP) testing is a concept originated in response to the rising number of alcohol and drug-related work incidents and injuries in industry. Such tests are designed to assess an employee's readiness to perform the job at the workplace. Gilliland and Schlegel (1993) defined RTP as "that state in which a person is prepared for a job, is capable of performing it, and is free of any transient risk factors that might influence performance". The first two elements of the definition assume that the employee is adequately trained, can meet the physical and emotional demands of the job, has adequate skills, and is appropriately motivated to perform the job. The primary focus of RTP testing lies in the third component of the definition. RTP testing is undertaken with the goal of assessing the transitory state of the employee, by determining if performance is altered due to the presence of risk factors.

In response to issues raised with traditional, and often invasive, techniques of biochemical and neurological testing, an alternate approach has been taken for readiness-to-perform testing. This approach utilizes performance on traditional neurological and cognitive tasks as the criterion to assess an employee's general level of work preparedness. The performance tests can be categorized according to the cognitive functions required to perform the tasks (e.g., motor, perceptual, or higher cognitive functions). Performance-based RTP tests are typically administered on a personal computer prior to the employee beginning the daily work shift. The computer can be used to automatically score performance and determine if the employee is ready to perform.

The development of performance-based RTP tests has proceeded rapidly and with little consideration to developing standards to categorize employee performance. Many of the current implementations of RTP tests use a seemingly arbitrary bound of 1.5 or 2.0 standard deviations from the employee's baseline to determine a lack of readiness. In other words, the range of acceptable performance is defined by the standard deviation measured across the baseline trials. This criterion has the potential to penalize consistent performers who have small standard deviations (i.e., narrower acceptable performance range) while allowing a larger acceptable performance range (in absolute terms) for highly irregular performers.

This paper examines the use of several statistical quality control techniques for evaluating individual performance on RTP tests. The techniques were applied to a database that represents performance under the influence of antihistamines. Each technique was applied to the database and a comparative evaluation of the approaches was accomplished. Performance of the techniques was assessed using actual risk factor presence as well as subjective judgments of out-of-control data points. The administration of an antihistamine as an actual risk factor provided a direct evaluation of the effectiveness for the various techniques for RTP testing. The subjective evaluation provided a measure of how well the techniques identified all "outlier" data, regardless of the nature of the irregularity.

METHOD

In order to effectively evaluate the merits of various techniques in detecting the presence of risk factors, performance data were needed. These data satisfied several requirements:

1. The data were relatively stable with minimal variation across trials, indicating that subjects were well trained and had accomplished a substantial portion of task learning,
2. there existed a sufficient number of baseline trials to "initialize" the statistical techniques, and
3. the data were reflective of performance under known, controlled, risk factor conditions.

The database selected for use in the analysis was collected as part of a contracted research effort for the Federal Aviation Administration (FAA) to address questions of reliability and validity associated with RTP testing (Gilliland and Schlegel, 1994). A subset of the original database was selected which included seven subjects, four tasks (yielding 17 criterion measures), and 48 trials. The final 18 trials were performed using a protocol to study antihistamine (4 mg. Chlor-trimeton™) effects. These trials were distributed across four weeks and included a total of six "refresher" trials to maintain performance proficiency and twelve experimental trials. Across the twelve experimental trials, both antihistamine and placebo doses were administered crossed with data collection on both day and night shifts. Three trials were collected in each of the four experimental sessions, defined by crossing drug dose with shift. Trials were conducted at one, five, and nine hours following dosing.

Once the subset database had been determined, the next step was to obtain an independent, subjective assessment of performance changes based on a "blind", visual pattern analysis of the data by a subject matter expert (SME) in the field of cognitive assessment. The SME was asked to provide the following two evaluations: determine the appropriate baseline data for initialization of the techniques, and identify any data trials that reflected irregular, or out-of-control, performance. The judgments of performance irregularity were used to evaluate each techniques' effectiveness in identifying any impaired performance.

In parallel with the independent SME performance assessment, the data were subjected to a variety of statistical quality control techniques for identifying out-of-control performance. The techniques that were evaluated included: 1) Shewhart charts (x-bar and s charts, p charts, or c charts, as dictated by the type of data; Montgomery, 1991; Wheeler and Chambers, 1992), 2) exponentially weighted moving average (EWMA) charts (Montgomery, 1991), 3) Cumulative Sum (CUSUM) charts (Breyfogle, 1992; Montgomery, 1991), and 4) a modified Shewhart chart based on the author's interpretation of currently applied RTP test methods. The charts were initialized using data points identified by the SME as representing stable, baseline performance (for all tasks, the last four data trials prior to antihistamine testing). Points prior to stabilization were excluded from the analyses due to the appearance of continued learning. The remaining trials (after data stabilization) were plotted on the control charts.

Each of the seventeen criterion measures were submitted to the relevant analysis techniques. In addition, the techniques were applied multiple times (for a total of 1715 charts) to examine the impact of modifying various chart parameters. Figures 1 and 2 illustrate the use of an EWMA chart and a CUSUM chart for evaluating one individual's reaction time on a Mathematical Processing task. All out-of-control points indicated by each chart analysis were recorded by trial for each set of chart parameters used. Actual risk factor condition (dose and shift) and SME evaluations of the data were also recorded by trial.

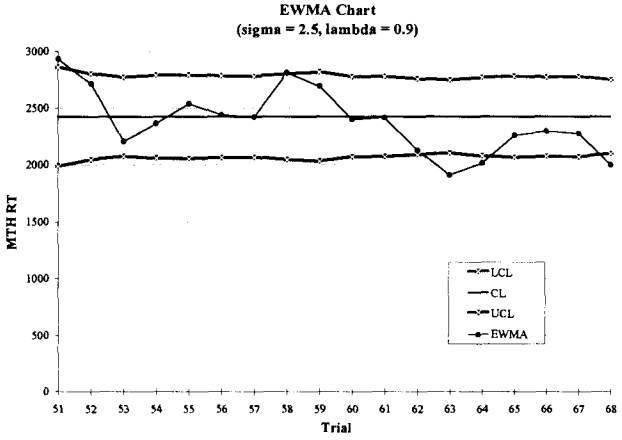

Figure 1. EWMA Chart of Single Subject Performance.

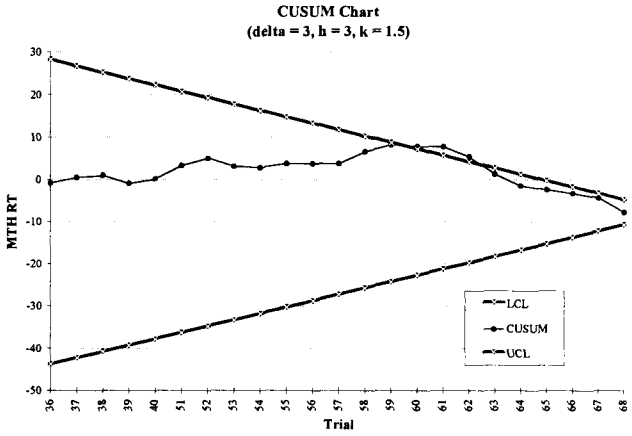

Figure 2. CUSUM Chart of Single Subject Performance.

Technique effectiveness was assessed using error measures analogous to those found in hypothesis testing epidemiology, and signal detection theory. Rates of hits, corrects, misses, and false alarms, and summary measures of sensitivity and specificity (Kennedy, Turnage, and Dunlap, 1992) were computed across subjects for each combination of criterion measure, technique, and chart parameter using either SME judgment or actual risk factor condition as "true" conditions. However, hit, miss, and sensitivity rates for actual risk factor conditions were summarized separately for the specific conditions of antihistamine-day shift (HD), antihistamine-night shift (HN), and placebo-night shift (PN).

The "refresher" and placebo-day shift (PD) trials were used to compute overall rates for correct, false alarm, and specificity. The resulting four sets of rates (by evaluation groups SME, HD, HN, and PN) were then used to determine an optimal parameter set for each combination of criterion measure, technique, and evaluation group, and the selected techniques (Shewhart, EWMA, CUSUM, and the modified Shewhart) were then analyzed using a series of pairwise tests of proportions. Any test with significantly fewer errors, was identified as being more effective for that criterion measure.

RESULTS

A subjective analysis was used to identify the optimal parameter set for each SQC technique and for each criterion measure based on the criteria of maximizing specificity while retaining high sensitivity. In many cases, the choice was obvious as both criteria were satisfied. In some instances, slightly lower specificity was accepted in exchange for large increases in sensitivity. If the trade-off between the criteria was more balanced, higher specificity was emphasized provided an adequate level of sensitivity was achieved. However, if the sensitivity was very low, decreased specificity was accepted in order to gain sensitivity. An additional criterion used when the trade-off was unbalanced was to select the technique that yielded substantially larger gains in hits as compared to increases in false alarms. Typically, the emphasis was placed on reducing false alarms and only small increases from the minimum were acceptable.

After an optimal parameter set had been selected for each technique and for the four evaluation groups, an attempt was made to identify which technique was most effective. Tests of hypotheses on two proportions (Hays, 1988; Hines and Montgomery, 1980) were used to discriminate between techniques. For each of the seventeen criterion measures and the four evaluation groups within each measure, both the sensitivity and the specificity indexes were subjected to pairwise tests of proportions. Each criterion measure had two sets of corresponding hypothesis tests for each of the four summary groups, one involving the specificity indexes and the other involving the sensitivity indexes. Each set of tests contained all pairwise combinations of the applicable techniques and the Bonferonni inequality (Hays, 1988) was used to control familywise alpha error rate. In most instances, the choice of the most effective technique for each criterion measure was obvious based on the results of the hypothesis tests.

The results of the tests are collectively summarized in Table 1 and Table 2 which present the most effective technique for each of the continuous and discrete criterion measures, respectively, and for each evaluation group. Each technique is identified by a two-character code. The first character represents the specific technique and the second character represents the parameter set for that technique. If no parameter set is indicated, then all parameter sets performed equally well. If no technique code is indicated, there were no significant differences among techniques. In addition, the maximum values of sensitivity and specificity obtained among the collection of effective techniques are provided. The final column of each table presents the recommended technique for the criterion measure based on the performance across all evaluation groups.

The continuous measures were successfully evaluated by the SQC control charts. As indicated by the summary column in Table 1, several techniques provided equivalent effectiveness for many of the continuous measures. Examination across the different evaluation groups suggests that continuous measures were best described by Shewhart and EWMA charts. For two continuous measures, there was no clearly superior technique. However, for five of the seven continuous measures for which a superior technique(s) was identified, CUSUM charts were equally effective. The modified Shewhart chart was effective for only three of the continuous criterion measures and not consistently effective across the different evaluation groups.

The SQC control charts were somewhat ineffective in clearly identifying performance changes for the discrete measures (see Table 2). The modified Shewhart charts were rarely able to attain even a modest level of sensitivity although their specificity was high. Typically, in the instances where standard Shewhart techniques were superior, it was due to a much higher level of sensitivity compared with the modified Shewhart charts.

It is interesting to note that, across all measures, the values of specificity were fairly high (0.67 to 1.00 with 71% above 0.90 and 84% above 0.80). This implies that the number of false alarm judgments is minimal relative to the number of correct judgments of no change in cognitive performance. Alternatively, the smallest values of sensitivity were very low (0.00 to 0.94 with 32% below 0.10 and 46% below 0.20) implying that more instances of impaired performance were missed than were hit. The strongest levels of sensitivity were obtained for the evaluation using the SME as the standard (0.67 to 0.94), implying high correspondence between the SME's identification and the technique's identification of hits and misses.

DISCUSSION

The goal of this research was to evaluate the use of statistical quality control charts as tools for identifying impaired performance of individuals in the RTP testing paradigm. Four control chart techniques were utilized and clear distinctions were evident between the techniques according to the type of data analyzed. Continuous measures were best evaluated with EWMA charts using a fairly large weighting factor (about 0.90). Shewhart charts were only moderately effective for the continuous data (about 50% of the measures). CUSUM charts were ineffective for most (almost 50%) continuous measures. Discrete data were well described using Shewhart p charts. Even when the modified Shewhart charts proved significantly better, the sensitivity was so low that the method would still be considered ineffective. As such, this research concludes that the standard deviation method

currently used in several RTP test schemes may have serious deficiencies. Alternate techniques, such as the control chart procedures recommended here, may better discriminate performance variation and lack of readiness to perform.

Table 1. Summary of Tests on Proportions for Continuous Measures.

Continuous Measure	RTP-SME	RTP-HD	RTP-HN	RTP-PN	Dominant Technique
1	S*3; E*7	S3; E*5, 8; V	S3; E*5, 8	S*3; E*5, 8; V	Shewhart/ EWMA
sensitivity	0.8667	0.2381	0.3333	0.0476	
specificity	1.0000	0.9592	0.9592	0.9592	
2	S2, 3; E*7	S*1, 4; E7; V	E7	S*1, 4; E7; V	Shewhart/ EWMA
sensitivity	0.7500	0.0476	0.0952	0.0952	
specificity	1.0000	0.8776	0.9388	0.8776	
3	E8	-	-	-	EWMA
sensitivity	0.6875	0.1429	0.1905	0.1905	
specificity	0.9620	0.9184	0.9184	0.9184	
4	E7	E*6, 12; C; V	E6, 12	E*6, 12; C; V	EWMA
sensitivity	0.9286	0.0476	0.1429	0.0476	
specificity	0.9881	0.9388	0.9388	0.9388	
5	E*5, 8	-	-	-	EWMA
sensitivity	0.9048	0.0952	0.1429	0.0952	
specificity	0.9778	0.9388	0.9388	0.9388	
6	S2; E*9	S*3, E12; C15	S3, E12; C*15	S3, E12; C*15	Shewhart/ EWMA
sensitivity	0.7647	0.3810	0.6190	0.4762	
specificity	0.8642	0.6735	0.6939	0.6939	
7	E12	S3; E6; C*13	S3; E*6; C13	S3; E6; C*13	EWMA
sensitivity	0.6750	0.7143	0.4286	0.5714	
specificity	0.9500	0.7347	0.8571	0.7347	
8	E12	S3; E*7; C15	S*3; E7; C15	S*3; E7; C15	EWMA
sensitivity	0.9429	0.4286	0.6190	0.3333	
specificity	0.9265	0.7347	0.7143	0.7143	
9	S2; E*6, 12	S2; E*9; C15	S*2; E9; C15	S2; E*9; C15	Shewhart/ EWMA
sensitivity	0.9286	0.6667	0.5238	0.3810	
specificity	0.8571	0.7143	0.6939	0.7143	

Legend: S - Shewhart x bar and s charts; E - EWMA charts; C - CUSUM charts; V - modified Shewhart charts. An asterisk (*) indicates the particular chart associated with the given sensitivity and specificity values.

Table 2. Summary of Tests on Proportions for Discrete Measures.

Discrete Measure	RTP-SME	RTP-HD	RTP-HN	RTP-PN	Dominant Technique
1	S4	V	V	V	Shewhart
sensitivity	0.9333	0.0000	0.0000	0.0000	
specificity	1.0000	1.0000	1.0000	1.0000	
2	-	-	-	-	
sensitivity	0.6667	0.0952	0.0000	0.0000	
specificity	1.0000	0.9184	0.9796	0.9796	
3	S4	V19, 20	V19, 20	S1, 3	Shewhart
sensitivity	0.8125	0.0000	0.0476	0.1905	
specificity	0.9818	0.9796	0.9796	0.8980	
4	S4	V19, 20	V19, 20	V19, 20	Shewhart
sensitivity	0.8571	0.0000	0.0476	0.0476	
specificity	1.0000	0.9796	0.9796	0.9796	
5	S4	S4	S4	V	Shewhart
sensitivity	0.6875	0.1429	0.2381	0.0000	
specificity	0.9813	0.9184	0.9184	1.0000	
6	S4	V19	V19	V19	Shewhart
sensitivity	0.9375	0.1905	0.0476	0.0952	
specificity	0.9222	0.9796	0.9796	0.9796	
7	-	-	-	-	
sensitivity	0.9200	0.1429	0.3333	0.2381	
specificity	0.9744	0.8980	0.8776	0.8980	
8	-	-	-	-	
sensitivity	0.9167	0.3333	0.4762	0.3333	
specificity	0.9770	0.8163	0.8163	0.8163	

Legend: S - Shewhart *p* or *c* chart; V - modified Shewhart charts.

REFERENCES

Breyfogle, F.W., III (1992). *Statistical methods for testing, development, and manufacturing*, New York: John Wiley and Sons, Inc.

Gilliland, K., and Schlegel, R.E. (1993). *Readiness-to-Perform Testing: A Critical Analysis of the Concept and Current Practices*. Final Report DOT/FAA/AM-93-13, Washington, DC: Office of Aviation Medicine, Federal Aviation Administration.

Gilliland, K., and Schlegel, R.E. (1994). *Development of a laboratory model of readiness-to-perform testing.* Paper presented at the Aerospace Medical Association 65th Annual Scientific Meeting, San Antonio, TX.

Kennedy, R.S., Turnage, J.J., and Dunlap, W.P. (1992). The use of dose equivalency as a risk assessment index in behavioral neurotoxicology. *Neurotoxicology and Teratology, 14*, 167-175.

Hays, W.L. (1988). *Statistics* (4th edition). Ft. Worth, TX: Holt, Rinehart, and Winston, Inc.

Hines, W.W., and Montgomery, D.C. (1980). *Probability and statistics in engineering and management science* (2nd edition). New York: John Wiley and Sons.

Montgomery, D.C. (1991). *Introduction to statistical quality control, 2nd Edition*, New York: John Wiley and Sons, Inc.

Shehab, R.L., Schlegel, R.E., and Gilliland, K. (1995). The use of statistical quality control techniques in readiness to perform testing. In *Proceedings of the Konz-Purswell Occupational Ergonomics Symposium* (pp. 109-114), Lubbock, TX: Texas Tech University.

Wheeler, D.J., and Chambers, D.S. (1992). *Understanding statistical process control*, Knoxville, TN: SPC Press.