Applying Quality Control Charts to the Analysis of Single-Subject Data Sequences

Randa L. Shehab and **Robert E. Schlegel,** University of Oklahoma, Norman, Oklahoma

Techniques from the field of quality control can be used to classify the quality of individual samples of physical or cognitive performance. After stable baselines have been established for an individual, deviations in performance can be evaluated using control charts. The effectiveness of this approach in evaluating cognitive performance was tested using databases collected under a variety of risk factors. The sensitivity and specificity characteristics of Shewhart, cumulative-sum (CUSUM), and exponentially weighted moving average (EWMA) control charts were determined for a total of 174 trials involving 10 participants and 23 cognitive performance assessment measures. The most effective technique in each case was typically a function of the specific performance measure and the type of performance change being evaluated. Sensitivity and specificity for the best techniques were as high as 100%. This study demonstrated the usefulness of quality control charts as a tool to evaluate individual participant performance over time. Actual or potential applications of this research include readiness-to-perform screening of industrial workers in order to improve the health and safety of the workforce.

INTRODUCTION

Background

The term *risk* refers to the potential for danger, loss, or injury. A *factor* is any condition that leads to a specific result. Taken collectively, a *risk factor* is any condition commonly considered to have the potential for negatively affecting or causing danger to an individual or within the individual's realm of responsibility. A multitude of risk factors exist that can be detrimental to human performance, and their effects on human performance have been studied extensively using descriptive and empirical research in both laboratory and applied environments. Risk factors such as alcohol (Gawron & Ranney, 1988; Maylor & Rabbitt, 1987), over-the-counter and illicit drugs (Hurst, 1976; Starmer, 1985), fatigue and sleep loss (Mertens & Collins, 1986; Steyvers, 1987), and various hostile environments have at some level been linked to either cognitive or physical performance decrements.

Exposure to these risk factors may occur off the job or result from hazards imposed by the nature of the job itself. It is common to think of alcohol and drug use when referring to such risk factors. However, fatigue, mental stress, illness, and even microgravity may present serious hazards to the employee and others affected by the employee's job performance. To minimize problems associated with on-the-job impairment, many employers have implemented screening programs designed to assess an employee's readiness to perform (RTP) at the workplace. Gilliland and Schlegel (1993) defined *readiness to perform* 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" (p. 3). RTP testing is undertaken with the goals of identifying changes in an individual's performance that may have been driven by exposure

Address correspondence to Randa L. Shehab, School of Industrial Engineering, University of Oklahoma, Norman, OK 73019; rlshehab@ou.edu. **HUMAN FACTORS,** Vol. 42, No. 4, Winter 2000, pp. 604–616.

to risk factors and determining the specific aspects of performance that are immediately affected.

Traditional methods of assessing readiness to perform focus on biochemical drug screening (Miller, 1994). Controversy surrounds drug screening, however, including issues concerning privacy, expense, data processing time, and specificity of results. An alternative method of assessing readiness to perform involves neurological evaluation. Although many of these tests are easy to administer and the results are immediately available (e.g., balance control and control of visual gaze), some of the more objective tests (e.g., electroencephalographic and pupillary responses) require extensive testing apparatus and training of the test administrator. Although biochemical and neurological techniques are often successful in detecting the use of chemical substances, they ignore other factors, such as fatigue and stress, that could potentially impair performance.

In an effort to avoid difficulties associated with biochemical and neurological testing, efforts have turned toward the development of performance-based assessment of RTP (Gilliland & Schlegel, 1995). These techniques use computer-based cognitive tasks administered before the employee begins work. Typical data collected from these tests usually include measures of reaction time and accuracy. A major challenge posed by the use of computer-based RTP testing is identifying an effective method of analyzing and interpreting performance data in order to make individualized judgments. Performance may vary considerably within and between individuals, and common parametric analyses based on pooling observations across participants often mask any relevant differences attributable to risk factors.

In the RTP scenario, each participant provides a single data set from each test by which performance is to be judged. The single-subject analysis approach often used with RTP implementations evaluates individual performance using seemingly arbitrary performance bounds based on the individual's variability across repeated sessions. The potential weakness of such an approach is that daily performance means are evaluated using session-to-session variability, and the influence of within-session variability is lost. More discriminating techniques for the analysis of individual participant data are critical to the successful implementation of performance-based RTP screening programs.

This paper addresses the development of effective analysis techniques for determining the presence of risk factors using performance-based measures. The proposed techniques capitalize on the within-session variability obtained by considering the multiple responses to stimuli within each session. The techniques were derived from statistical quality control (SQC) and modified to fit the RTP paradigm. Data collected from 10 participants, 174 trials, and 23 performance measures were analyzed using 18 variations of three different quality control charts. Additionally, the data represent performance under space microgravity or antihistamine conditions. The control chart techniques were evaluated in terms of their ability to correctly diagnose the presence of risk factors while minimizing the occurrence of false alarms.

RTP Scoring Techniques

Three existing techniques for evaluating RTP task performance were identified. The most common procedure is the use of individual performance variability to determine out-of-bounds performance (Miller, Kim, & Parseghian, 1995; O'Donnell, 1991). A predetermined number of trials is used to estimate parameters of the sampling distribution of the performance measure, which is typically the mean of the responses to multiple stimuli. The overall mean is used as the target performance level, and an arbitrary multiple of the standard deviation forms the performance boundaries. Performance outside the boundaries identifies individuals who are not ready to perform.

Kennedy, Turnage, and Dunlap (1993) and Kennedy, Turnage, and Jones (1995) utilized *performance loss,* a relative change from baseline, to evaluate performance on RTP tasks. This technique uses arbitrary loss bounds. Kennedy et al. (1993) selected 10% as the loss bound for performance evaluated with their automated performance test system, and a series of multiple cutoffs was used to reduce unacceptably high false alarm rates (Kennedy et al., 1995).

A third technique for evaluating RTP task performance uses neural networks (O'Donnell, Fix, & Morton, 1995). The neural network is trained to recognize unimpaired performance on a given task and computes an associated activity measure. When future network activity substantially deviates (based on an arbitrary cutoff) from the unimpaired activity level, the participant is considered impaired.

Statistical Quality Control Analogy

Statistical quality control (SQC) techniques were developed to determine process capabilities and evaluate process performance. Process variability in SQC is distinguished as common-cause versus special-cause variability. *Common-cause variability* represents inherent process variation beyond the control of the worker. *Special-cause* (or assignable-cause) *variability* results from sources external to the process and is manifested as an abnormal deviation on a single trial or a sustained shift (instantaneous or developing trend) in the process. In SQC, the organization is responsible for identifying and eliminating special-cause variability.

RTP evaluations are directly analogous to SQC evaluations. In RTP, the process is the performance of human tasks. Such performance is subject to inherent human (i.e., common-cause) variability. Special-cause variability is analogous to risk factors that directly affect worker performance. If only common-cause variability exists, performance is assumed to be stable or asymptotic for that individual, and no further learning is evident. This is in contrast to differential stability, which allows continued improvement at the same rate for all members of a participant group. The goal of an RTP evaluation is to individually differentiate between stable performance and performance changes caused by the presence of risk factors. Affected workers are then removed or reassigned, or job requirements are modified.

The traditional SQC tool to evaluate process performance is the Shewhart control chart. Shewhart charts evaluate a process using control limits defined in terms of the process standard deviation. As compared with current RTP techniques, which use the standard deviation across trials, Shewhart control limits are determined using the standard deviation within each sample (i.e., trial) of data. Shewhart charts also provide a two-level assessment of the process. First, the variation of the process is evaluated with either a range (R) or standard deviation (SD) chart. These charts plot the relevant estimate of variability. The process variability is considered stable from trial to trial if all plotted points fall within the control limits. Then the process mean is examined using an x-bar chart, which plots the sample (or subgroup) mean for each trial and identifies points that exceed the control limits. Shewhart charts have been shown to be effective for identifying process shifts as small as 1.5 standard deviations (Montgomery, 1997) and have been developed for continuous data (e.g., reaction times) and for data based on discrete events (e.g., percentage correct).

Two alternatives to the Shewhart control chart are often more effective when the detection of small (sustained) shifts is of interest. The cumulative-sum (CUSUM) control chart evaluates accumulated deviations across samples between the mean performance and the desired performance level to identify process shifts (Breyfogle, 1992; Montgomery, 1997). Performance changes are manifested as a change in the slope of the cumulative sum. This slope is initially zero for a stable process centered at the target mean (Vardeman & Jobe, 1999). Because consecutive cumulative sums are correlated, conventional (horizontal) control limits are replaced with a V-mask defined by error probabilities and the magnitude of the desired shift to be detected. The V-mask serves to identify points outside the desired range of performance and can be described as a horizontally rotated V shape with the vertex positioned a fixed distance ahead of the current data point and the arms extending back toward the origin. When the plotted data are overlaid with the V-mask, any previous sum falling outside the mask indicates that the process shifted at some point prior to the current sample.

Alternatively, a tabular monitoring approach can be used. CUSUM charts are often used for individual observations ($n = 1$). The practice of standardizing the performance variable prior to plotting simplifies the selection of parameters and enables a better understanding of the CUSUM method.

Exponentially weighted moving average (EWMA) charts are a third genre of SQC charts that utilize smoothed data. EWMA charts plot a weighted average of subgroup (i.e., trial) means using an exponential weighting that decreases geometrically with age. Control limits are used with EWMA charts and are computed as a function of the standard deviation and the exponential weighting factor. The weighted value is judged against these control limits. For a weighting factor of 1, the EWMA data series corresponds to the raw data series used in the Shewhart chart (Vardeman & Jobe, 1999). Because a measure of variability within each trial is not required for the calculations, EWMA charts can easily be used for individual observations ($n = 1$).

Of the three SQC techniques, Shewhart charts are considered best for identifying abnormal deviation on a single trial and for detecting large shifts in the process. CUSUM and EWMA charts perform better for identifying smaller process shifts and trends because they utilize the performance data collected across multiple trials. EWMA charts perform considerably better than CUSUM charts for detecting large process shifts (Montgomery, 1997).

METHODOLOGY

To evaluate the relative merits of various SQC techniques in assessing the presence of risk factors, performance data were needed. The data used in the analysis were obtained from two previously collected databases. Each database satisfied several criteria necessary for successful implementation of the SQC techniques. First, the data exhibited a relatively stable baseline such that participant performance without risk factor exposure demonstrated little trial-to-trial variation and indicated minimal continued learning. Second, the databases provided a sufficient number of trials at the baseline level of performance to adequately initialize each SQC technique. The final critical characteristic of each database was the inclusion of trials collected under risk factor conditions. These trials were subjected to analysis such that the techniques could be compared for their ability to identify the stressor trials.

The two databases analyzed in this study were developed using cognitive performance assessment batteries. The first database was collected under contract to the National Aeronautics and Space Administration (NASA; Schiflett, Eddy, Schlegel, French, & Shehab, 1995) and provides performance data collected during the space flight of three male astronauts. The second database provides extended performance data on 16 male college students subjected to antihistamine doses and sleep loss and was collected under contract to the Federal Aviation Administration (Gilliland & Schlegel, 1994). Subsets of each database were selected such that a diverse set of cognitive and psychomotor performance tasks was assembled. The tasks included critical tracking, spatial matrix, Sternberg memory search, continuous recognition memory, attention switching (which combines mathematical processing and manikin tasks), and a dual task that combines memory search with tracking at both individualized and group levels of task difficulty. Table 1 summarizes the tasks and their respective criterion measures. A complete description of each of the tasks and measures can be found in Schlegel, Shehab, Gilliland, Eddy, and Schiflett (1995). Performance data for all tasks except dual task group (DULG) were included from Database 1, whereas data for only critical tracking (TRK), dual task individual (DULI), DULG, and attention switching were included from Database 2.

The stressor conditions differed for the two databases. Database 1 was developed to study the effects of the space environment, particularly microgravity, on human cognitive performance (Schiflett et al., 1995). Three astronauts performed 24 preflight trials, 13 in-flight trials, and 3 postflight trials over a period of approximately 8 weeks. For Database 2, which was developed to study the effects of antihistamines and sleep loss on human cognitive performance (Gilliland & Schlegel, 1994), 16 participants performed 30 practice trials before beginning the risk factor investigation period following a delay of several weeks. This risk factor testing was spread across five weekends with an idle weekend between the second and third weekends. Participants performed two refresher trials during the weekdays before the first and third weekends of testing and one

TABLE 1: Performance Tasks and Measures

Task	Task Code	Lambda (LM)	Correct Reaction Time (RT)	Percentage Incorrect (PI)	Root Mean Square Error (RMS)	Control Losses (CL)
Critical tracking	TRK	X				
Spatial matrix	MTX		X	X		
Sternberg memory search	STN		X	X		
Continuous recognition	CRC		X	X		
Attention switching						
Manikin	MAN		X	X		
Math processing	MTH		X	X		
Transition trials	MANX/MTHX		X	X		
Dual task	DUL					
Memory search			X	X		
Unstable tracking					X	X

refresher trial during the weekdays before the second and fourth weekends of testing. Across the four sessions of risk factor testing, participants were assigned at random combinations of antihistamine dose (4 mg Chlor-Trimeton™ or placebo) and work shift (day or night). Seven participants were chosen at random from Database 2 for inclusion in the current analysis.

Once the databases had been identified, the data were subjected to an independent assessment of performance changes by a subject matter expert (SME) in the field of human cognitive performance. This assessment was helpful because a participant's performance may not always correlate with the presence or absence of a risk factor. For example, exposure to a risk factor may not always be reflected by a change in performance; alternatively, performance may indicate the presence of an external risk factor not specified in the experimental protocol. The assessment identified actual performance impairment based on a visual pattern analysis of the data. The analysis was blind in that the evaluator was unaware of the experimental conditions associated with each data point. In addition, intrarater reliability was evaluated by comparing the judgments with a second set of judgments performed by the same SME 6 weeks later. The intrarater reliability was high, with 91% of the 912 judgments in agreement between rating sessions.

The methodology by which the SQC techniques were evaluated is illustrated in Figure 1. The statistical techniques proposed to identify risk-factor-induced performance changes were evaluated by subjecting the various measures in each database to each of the analysis techniques. Three techniques were evaluated: (a) Shewhart charts (either x-bar and s charts or a p chart), (b) CUSUM charts, and (c) EWMA charts.

Each technique was implemented using a variety of parameter configurations, which served to provide an opportunity for a broad range of chart performance. The parameters for the Shewhart charts define the width of the control limits and were specified as multiples of the standard deviation (σ) of the charted parameter or as the desired level of the Type I error probability (α). The EWMA charts use similar parameters to define the control limits but also include a parameter (λ) to define the weight allocation between the current data point and previous data points. CUSUM parameters are used to define the V-mask shape and position and can be specified with one of two alternative parameter sets. The first set specifies the Type I error probability (α), the Type II error probability (β), and the shift in the process mean (δ) to be detected. The alternative parameter set specifies the decision limits (h) about the current plotted

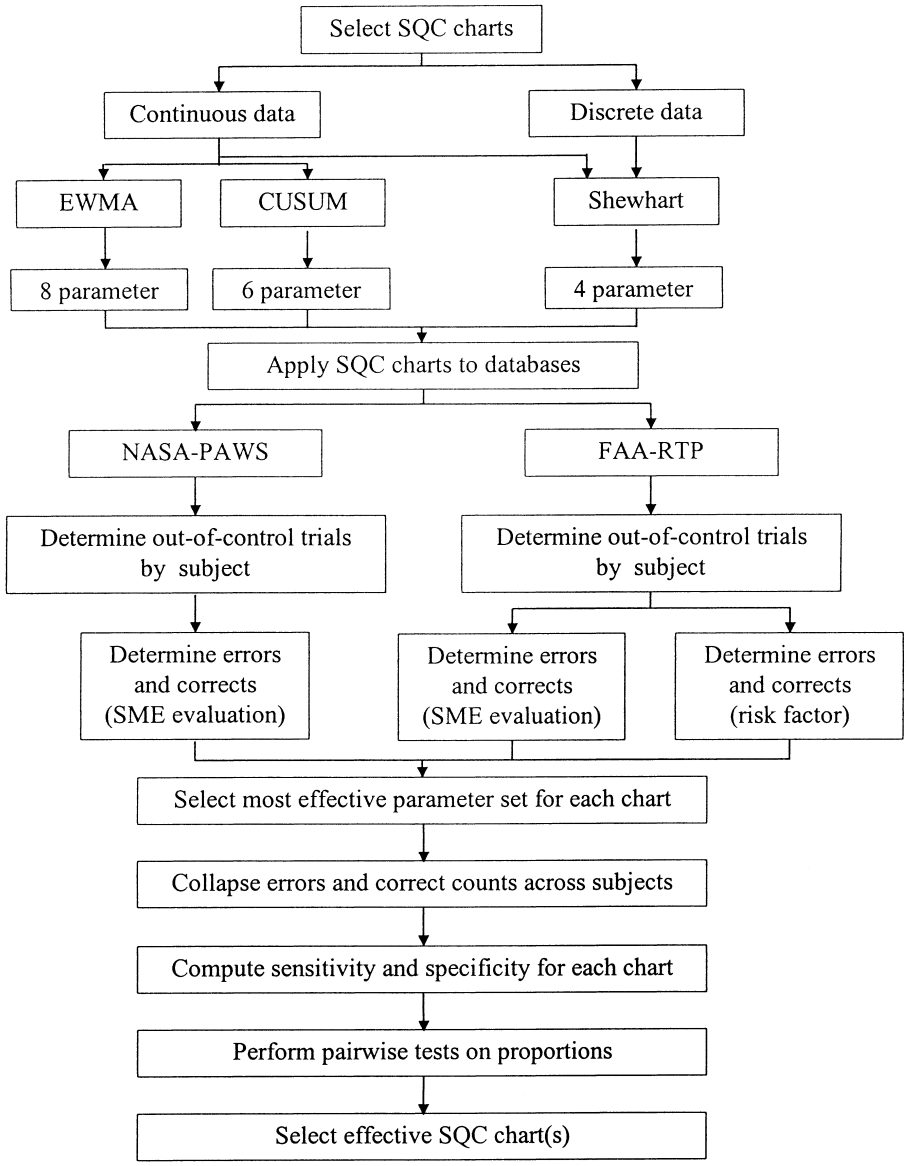

Figure 1. Methodology for evaluation of SQC charts for each performance measure.

value and the slope of the arms of the V-mask (k) in terms of multiples of the standard deviation.

Baseline data were used to initialize the chart parameters. It is generally recommended that 20 samples be used to initialize Shewhart charts (Wheeler & Chambers, 1992). Similar recommendations might be made for the other charts when their control limits are defined based on the standard deviation. However, these samples would typically each contain 4 to 6 data points (for a total of approximately 100 measurements), compared with the 30 to 150 data points per sample from the cognitive databases. Although each performance data set contained at least 24 trials prior to risk factor exposure, it was observed that the early trials exhibited a substantial degree of instability caused by continued participant learning of the tasks. The best estimates of stable means and standard deviations were provided by the last three practice sessions for Database 1 and the

last five practice sessions for Database 2. For the discrete stimulus tasks, these parameter estimates are thus based on 100 to 750 individual responses per participant.

The remaining trials in each data set, including data obtained under risk factor conditions, were analyzed using each technique. Control charts were generated for each of the techniques using the various combinations of participant, task, performance measure, and parameter configuration. Each of the generated control charts was analyzed according to the appropriate rules for identifying an out-of-control process point. The CUSUM charts were evaluated using V-masks, and for the other charts we used their specific application of control limits.

After all data sets had been charted, the performance of each technique was evaluated. The results of the charting analyses were represented as correct (correct rejections and hits) and incorrect (false alarms and misses) classifications of impairment. The basis of comparison was either the judgment of the SME or the presence/absence of a risk factor. These values were then summarized across participants for both databases and incorporated into measures of specificity and sensitivity (Kennedy, Turnage, & Dunlap, 1992). Specificity describes the classification of judgments for the unimpaired population and is the ratio of the number of trials correctly classified as unimpaired divided by the total number of unimpaired trials. Sensitivity describes the accuracy of the classification of judgments for the impaired population and is the ratio of trials correctly classified as impaired divided by the total number of impaired trials.

In the context of RTP, equivalent values of specificity and sensitivity differentially affect the number of classification errors made. For a sufficiently low population impairment rate, a given change in specificity would result in a substantially greater change in the overall error rate than would the same change in sensitivity (i.e., false alarms vs. misses, respectively). As a result, Kennedy et al. (1992) suggested that in RTP applications, specificity is more critical than sensitivity because even a low false alarm rate can result in removing or reassigning a large number of unimpaired employees.

Before comparing techniques, it was necessary to determine the most effective parameter configuration for each technique. This selection was based on the dual criteria of maximizing specificity while retaining high sensitivity. In many instances the optimal parameter configuration was obvious because one configuration exhibited substantially higher values of specificity and sensitivity. Some selections involved a trade-off such that a lower specificity was accepted in exchange for a substantial improvement in sensitivity. However, minimizing the false alarm rate was always emphasized, and only small deviations from that minimum were tolerated.

Techniques were then compared using the optimal parameter configuration identified for each technique. Sensitivity and specificity indexes from these techniques were compared using tests of hypotheses on two proportions. Three charting techniques were applied to the reaction time (RT), root mean square error (RMS), and lambda (LM) measures, so three paired tests were conducted. The Bonferroni inequality was applied to the analyses to control the family-wise Type I error rate (Hays, 1988). The percentage incorrect (PI) and control losses (CL) measures were examined with only the Shewhart chart, and no paired test was necessary. If the results of the tests on proportions were significant, an optimal control chart technique was selected for each performance measure.

RESULTS

Database 1

The evaluation of chart effectiveness for the NASA astronaut performance data was based on the SME judgments. These judgments were considered to reflect the true impact of microgravity on participant performance. In other words, the SME judgments were used as the basis for evaluating whether performance was impaired relative to baseline standards. Frequency counts of errors and correct judgments were determined by comparing control chart judgments with SME judgments.

A control chart was developed for each performance measure and participant using each

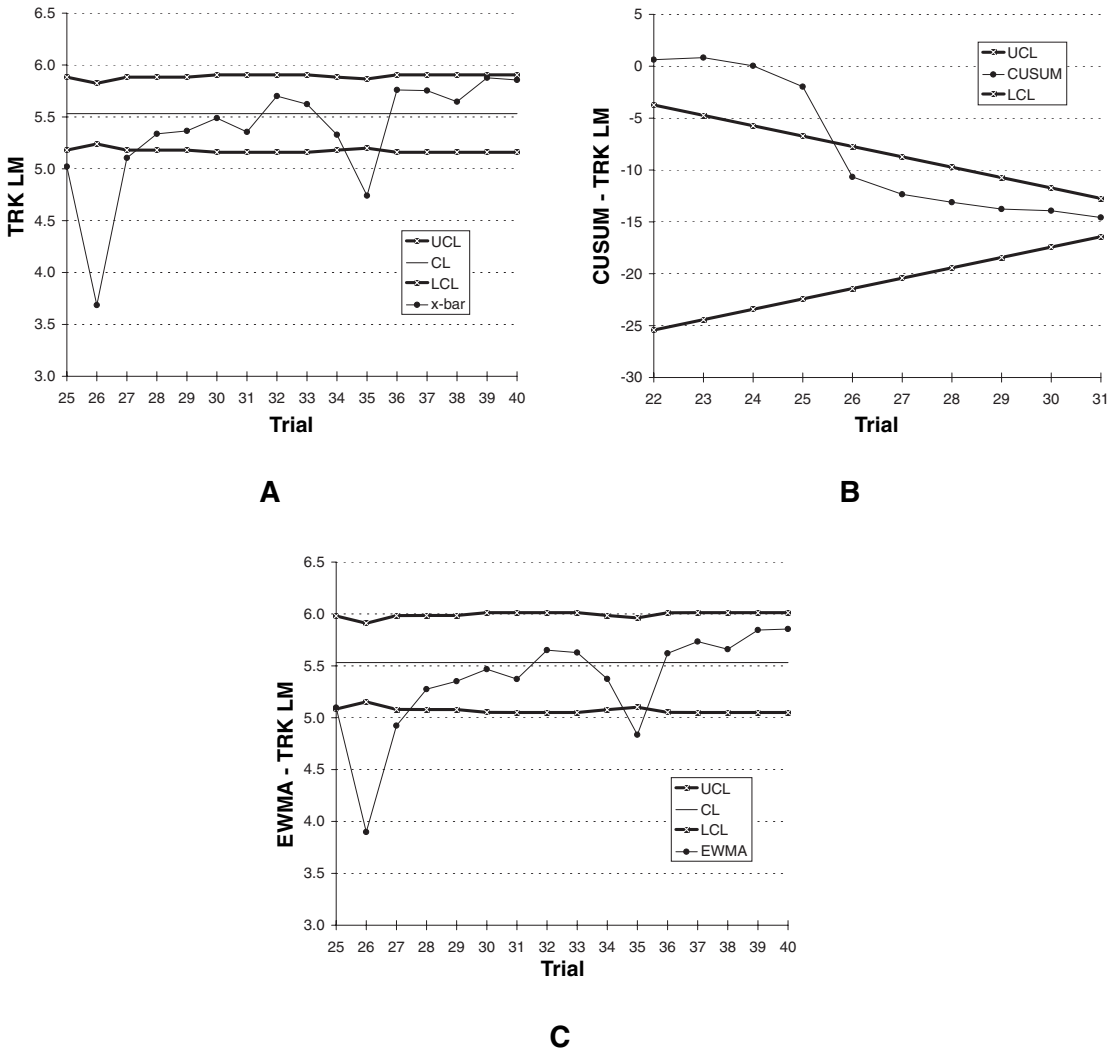

Figure 2. SQC charts for tracking task mean lambda (Database 1, Participant 002): (a) Shewhart chart (±2 σ), (b) standardized CUSUM chart (δ = 3, h = 1.5, k = .75), and (c) EWMA chart (±3 σ, λ = .85).

of the techniques with the various parameter sets. This process yielded 648 control charts for Database 1. Three SQC charts are shown in Figure 2 for the critical tracking task mean lambda performance measure for a single participant. For this measure, low values represent impaired performance.

The Shewhart chart in Figure 2a shows that Trials 25, 26, 27, and 35 are "out of control." The same data (standardized and accumulated) are presented in a CUSUM chart (Figure 2b). The chart illustrates the evaluation with the V-mask anchored at Trial 31 and clearly shows the performance change that occurred between Trials 25 and 26. The EWMA chart (Figure 2c) indicates Trials 26, 27, and 35 as out of control.

The SME judged trials 25, 26, 27, and 35 to be out of control, directly corresponding to the trials identified by the Shewhart technique. Thus for the Shewhart chart with the specified parameter set, the values for sensitivity and specificity are both 1.0. In general, the CUSUM chart identified a large number of trials as out of control, thus yielding strong sensitivity (1.0) but poor specificity (often 0.0). The corresponding values of sensitivity and specificity were .75 and 1.0, respectively, for the EWMA chart.

After evaluating the performance measure for each participant with all control chart techniques, the frequency counts were collapsed across participants in order to compute overall indexes of sensitivity and specificity. Using these overall indexes, the most effective parameter set for each technique was selected. In the case of conflicting results, a subjective determination was made using the criteria of maximizing specificity while retaining acceptable sensitivity. These selected techniques were then subjected to paired tests of proportions on the indexes of sensitivity and specificity in order to identify which of these techniques (if any) was most effective for the performance measure in question. Finally, each technique determined to be most effective was submitted to a test of a single proportion equal to a target value of .5 to determine whether the test performed better than chance. A directional alternative hypothesis (greater than chance) was employed.

Database 2

Evaluation of Database 2 was performed using two different foundations. First, control chart performance was compared against SME judgments. This analysis was identical in procedure to that used for Database 1. The second foundation used the actual risk factor condition as the basis for defining stressor exposure. However, the analysis procedure was the same as that previously employed. To avoid confounding the experimental variables, each combination of antihistamine level and work shift was considered as a separate risk factor condition. In other words, the risk factors were considered to be antihistamine-day shift (HD), antihistamine-night shift (HN), and placebo-night shift (PN). These trials were used to compute hit and miss frequencies and sensitivity rates for each risk factor condition. The "refresher" trials and the placebo-day shift (PD) trials were used in the computation of a single set of correct and false alarm frequencies and the overall specificity rate. Overall, 1358 charts were generated from Database 2.

Figure 3 presents the Shewhart *p* chart using data for a single participant taken from the manikin task percentage incorrect measure. Examination of the chart indicates that Trials 52, 53, 54, 59, 60, and 67 are out of control. Substantial correspondence was obtained between the *p* chart and SME judgments, with values for specificity and sensitivity as high as 1.0. The ability of the chart to identify actual risk factor exposure was poor. The sensitivity for the Shewhart *p* charts ranged from 0.0 (for placebo-night) to .67 (for antihistamine-day), and specificity never exceeded .71.

After participant performance was evaluated individually, the frequency counts were collapsed across participants, and two sets of indexes were computed: one for SME and one for risk factor. These overall indexes were used to determine the optimal parameter set for each technique. Using the indexes from the optimal parameter sets, tests of proportions were conducted to select the best technique or techniques and to determine whether the selected technique performed better than chance. The SME-based evaluation was tested separately from the risk factor evaluations. Each risk factor condition test on specificity rates was based on the same refresher and PD trials. However, the sensitivity tests differed because they were based on specific experimental conditions (i.e., HD, HN, PN).

Summary across Databases

Tests of proportions were used to identify the optimal technique or techniques for different performance metrics to investigate individual participant performance. From Database 1,

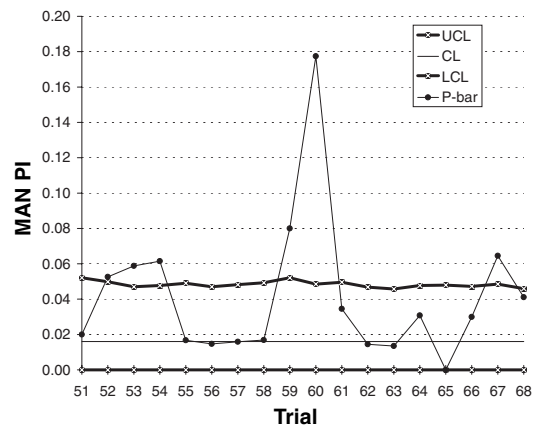

Figure 3. SQC *p* chart (±2 σ) for manikin percentage incorrect (Database 2, Participant 002).

each of the 10 continuous performance measures was effectively assessed by at least one Shewhart chart or one EWMA chart. As expected, tighter chart boundaries were typically more successful at detecting out-of-control data points. For discrete measures (e.g., PI) from Database 1, *p* charts with tighter boundaries were the most effective.

Similar results were found for the evaluation of Database 2 against the SME judgments. However, the evaluation of these data against the various risk factor conditions yielded slightly different results. Shewhart and EWMA charts performed significantly better than other charts for seven of the continuous measures. However, CUSUM charts also performed well on five of the seven measures.

Beyond identifying a more effective technique, it is important that the technique provide strong values of sensitivity and specificity. In general, values of specificity across the more effective techniques were fairly high (from .6735 to 1.000, with a mean value of .9263), implying effective control of false alarms. The task consistently yielding the lowest specificity was DUL for both the RT and CL measures. Overall, the measures of sensitivity were somewhat lower (from 0.0000 to 1.000 with a mean value of .4907), indicating that even the more effective techniques missed out-of-control data points on occasion. Although Kennedy et al. (1995) diminished the relative importance of sensitivity, some of the values are so low that the techniques may miss most of the impaired participants. However, if the data are examined using only the SME as the frame of reference, the mean values of specificity (.9525) and sensitivity (.7789) increase substantially.

Table 2 presents a summary of the SQC methods identified from the tests of proportions for tasks and measures common to both databases. The results for each data set and across data sets indicate which chart provided better identification of cognitive performance changes. When the test of proportions revealed no significant differences among multiple charts, those charts were indicated. Along with identifying the most successful SQC method(s), Table 2 presents the maximum values of sensitivity and specificity (corresponding to the identified method).

CONCLUSIONS

The goal of this research was to investigate the effectiveness of using three statistical quality control methods for identifying risk-factor-impaired cognitive performance. Of the three quality control chart techniques examined, clear discriminations were made among the techniques appropriate for different types of data. Continuous performance measures (e.g., reaction time, mean lambda, RMS error) were best evaluated with exponentially weighted moving average (EWMA) charts having a large weighting factor (about .90). Shewhart charts were moderately effective for these data, but cumulative-sum (CUSUM) charts were grossly ineffective. However, percentage-incorrect data were well described using Shewhart *p* charts.

We believe that the differences in effectiveness among the charts were attributable primarily to the type of performance change detected rather than specific characteristics of the task measures. Exposure to a risk factor for a single trial (e.g., gravity transitions or antihistamine dosing) produces a single out-of-control point, which is most effectively detected with a Shewhart chart. If the magnitude of the abnormal deviation is sufficient, the point will also be identified by an EWMA chart with a large weighting factor (λ) and, less often, by the CUSUM chart. However, risk factors that develop effects that are sustained across several trials (e.g., fatigue from sustained operations or sleep loss) produce an accumulated performance shift that is more effectively detected by the EWMA and CUSUM charts.

A closer review of the CUSUM (and EWMA) chart performance in this study revealed greater effectiveness across trials that reflected fatigue development (e.g., later sessions of the astronaut missions). The relatively low specificity for the CUSUM charts was primarily a result of the parameter sets selected and the implementation approach used. Performance of these charts in detecting sustained performance shifts or trends may be improved through the use of a different implementation method that reduces the number of false alarms, but it is highly unlikely that CUSUM charts can effectively detect one-time performance deviations.

An interesting result of this research was

TABLE 2: Summary of Tests on Proportions for SQC Charts

Measure	Database 1 SME	Database 2 SME	Database 2 HD	Database 2 HN	Database 2 PN	Effective Charts
TRK LM	S, E	S, E	S, E	S, E	S, E	Shewhart/
Sensitivity	.9474	.8667	.2381	.3333	.0476	EWMA
Specificity	1.0000	1.0000	.9592	.9592	.9592	
MAN RT	S, E, C	S, E	S, E	E	S, E	Shewhart/
Sensitivity	.8571	.7500	.0476	.0952	.0952	EWMA
Specificity	.9756	1.0000	.8776	.9388	.8776	
MANX RT	S, E, C	E	S, E, C	S, E, C	S, E, C	EWMA
Sensitivity	1.0000	.6875	.1429	.1905	.1905	
Specificity	1.0000	.9620	.9184	.9184	.9184	
MAN PI	S	S	S	S	S	Shewart
Sensitivity	1.0000	.9333	.0476	.0476	.0476	
Specificity	.6744	1.0000	.9388	.9388	.9388	
MANX PI	S	S	S	S	S	Shewart
Sensitivity	1.0000	.6667	.0952	.0000	.0000	
Specificity	.8182	1.0000	.9184	.9796	.9796	
MTH RT	S, E	E	E, C	E	E, C	EWMA
Sensitivity	.5455	.9286	.0476	.1429	.0476	
Specificity	.9730	.9881	.9388	.9388	.9388	
MTHX RT	S, E, C	E	S, E, C	S, E, C	S, E, C	EWMA
Sensitivity	.3000	.9048	.0952	.1429	.0952	
Specificity	1.0000	.9778	.9388	.9388	.9388	
MTH PI	S	S	S	S	S	Shewart
Sensitivity	.4286	.8125	.0000	.0952	.1905	
Specificity	.9756	.9818	.8980	.8980	.8980	
MTHX PI	S	S	S	S	S	Shewart
Sensitivity	.2857	.8571	.0000	.1429	.0952	
Specificity	.8780	1.0000	.9184	.8980	.8980	
DULI RT	S, E, C	S, E	S, E, C	S, E, C	S, E, C	Shewart/
Sensitivity	.8333	.7647	.3810	.6190	.4762	EWMA
Specificity	.8889	.8642	.6735	.6939	.6939	
DULI PI	S	S	S	S	S	Shewart
Sensitivity	.8750	.6875	.1429	.2381	.0476	
Specificity	.9750	.9813	.9184	.9184	.9184	
DULI RMS	S, E	E	S, E, C	S, E, C	S, E, C	EWMA
Sensitivity	.9231	.9429	.4286	.6190	.3333	
Specificity	.9429	.9265	.7347	.7143	.7143	
DULI CL	S	S	S	S	S	Shewart
Sensitivity	.9167	.9200	.1429	.3333	.2381	
Specificity	.8888	.9744	.8980	.8776	.8980	

Note. Chart techniques are denoted by S = Shewhart, C = CUSUM, and E = EWMA. The basis of comparison is SME = subject matter expert; HD = antihistamine, day shift; HN = antihistamine, night shift; PN = placebo, night shift. Values for sensitivity and specificity are shown for the chart with optimal performance on that task measure. See Table 1 for explanation of other codes.

the marked difference in sensitivity and specificity between the SME and risk factor evaluations. The SME-based evaluation yielded higher values of sensitivity and specificity than did the analysis of data in reference to risk factor condition. This phenomenon is attributed to the fact that people have individual susceptibilities to the experimental risk factor doses, and they may or may not exhibit a concomitant change in performance. The SME analysis incorporates this individual difference by considering the data participant by participant rather than based on assumed impairment caused by risk factor dosing.

To strengthen the conclusions drawn from this study, expanded databases should be examined with these techniques. For the tasks common to both databases, a total of 10 participants were included in the analysis. For tasks unique to each database, this number was substantially fewer. Additional participants would strengthen the statistical conclusions and would provide greater confidence in the results. More risk factors need to be examined to determine whether the patterns indicated for continuous and discrete data are replicated. Specifically, risk factors anticipated in a typical work environment should be investigated. Short-term space travel as a risk factor applies to a very limited population, and antihistamine represents, typically, a mild risk factor within the work environment. The inclusion of additional relevant risk factors and higher doses would also enhance the generalizability of the results.

The analyses described in this paper demonstrated the effectiveness of statistical quality control techniques for determining the quality of individual human performance. Although this paper examined cognitive performance, we believe this approach would be equally useful for analyzing other metrics of human performance. The primary requirement for using control charts to identify significant deviations in single-subject performance is the attainment of stable baseline performance indicating a minimal level of continued learning. Several trials of stable performance are needed to initialize the chart parameters. Performance trials collected beyond these initialization trials can then be monitored via SQC control charts. After each performance trial, the data can be plotted on the chart and evaluated against the performance boundaries specified. If performance exceeds the boundaries (in the direction of degraded performance), then performance is considered to be out of control and the appropriate action can be taken.

The design of SQC charts to meet specific detection goals and to achieve specified error probabilities can be readily accomplished using any of the texts cited in the references. Implementation of SQC charts for evaluating individual human performance deviations can be completed using a number of statistical packages such as SAS (SAS Institute, Cary, NC) or Statgraphics (Manugistics, Inc., Rockville, MD). For each test session involving multiple stimuli, the data requirements consist of the number of stimuli along with the mean and standard deviation of the response measure or measures for the session. Although this paper has presented support for the use of SQC techniques in RTP scoring, other issues exist concerning the implementation of an RTP program. These issues are described in detail in Gilliland and Schlegel (1993, 1995) and include hardware and software requirements, data processing, nonproductive test time, and a variety of theoretical validity issues.

SQC charts are a common tool used by industry practitioners to monitor process performance. By adapting these familiar methods to the monitoring of human performance in industry, the process of screening workers for the presence of risk factors is easily understood and more readily implemented. The use of SQC techniques in conjunction with RTP testing can improve the health and safety of the workforce.

REFERENCES

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

Gawron, V. J., & Ranney, T. A. (1988). The effects of alcohol dosing on driving performance on a closed course and in a driving simulator. *Ergonomics, 31,* 1219–1244.

Gilliland, K., & 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: Federal Aviation Administration, Office of Aviation Medicine.

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

Gilliland, K., & Schlegel, R. E. (1995). Readiness-to-perform testing and the worker. *Ergonomics in Design, 3*(1), 14–19.
Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart, and Winston.
Hurst, P. M. (1976). Amphetamines and driving behavior. *Accident Analysis and Prevention, 8,* 9–13.
Kennedy, R. S., Turnage, J. J., & Dunlap, W. P. (1992). The use of dose equivalency as a risk assessment index in behavioral neurotoxicology. *Neurotoxicology and Teratology, 14,* 167–175.
Kennedy, R. S., Turnage, J. J., & Dunlap, W. P. (1993). Diagnosis of alcohol intoxication: Effectiveness of cognitive and neurovestibular field sobriety tests. In *Proceedings of the Human Factors Society 37th Annual Meeting* (pp. 964–967). Santa Monica, CA: Human Factors and Ergonomics Society.
Kennedy, R. S., Turnage, J. J., & Jones, M. B. (1995, May). *Fitness-for-duty using multiple cut-offs to reduce false positives and retain sensitivity in computerized performance testing.* Paper presented at the Aerospace Medical Association 66th Annual Scientific Meeting, Anaheim, CA.
Maylor, E. A., & Rabbitt, P. M. A. (1987). Effects of alcohol and practice on choice reaction time. *Perception and Psychophysics, 42,* 465–475.
Mertens, H. W., & Collins, W. E. (1986). The effects of age, sleep deprivation, and altitude on complex performance. *Human Factors, 28,* 541–551.
Miller, J. C. (1994). *Industrial fitness-for-duty testing* (Final Report DTFH61-93-C-00088). Washington, DC: Federal Highway Administration, Trucking Research Institute of the American Trucking Association.
Miller, J. C., Kim, H. T., & Parseghian, Z. (1995, May). *Feasibility of in-cab fitness-for-duty testing of commercial drivers.* Paper presented at the Aerospace Medical Association 66th Annual Scientific Meeting, Anaheim, CA.
Montgomery, D. C. (1997). *Introduction to statistical quality control* (3rd ed.). New York: Wiley.
O'Donnell, R. D. (1991). *Scientific validation of the NOVASCAN™ tests: Theoretical basis and initial validation studies.* (Available from Nova Technology, Inc., 19460 Shenango Drive, Tarzana, CA 91356)
O'Donnell, R. D., Fix, E., & Morton, P. (1995, May). *Non-linear analysis techniques in readiness-to-perform testing.* Paper presented at the Aerospace Medical Association 66th Annual Scientific Meeting, Anaheim, CA.
Schiflett, S. G., Eddy, D. R., Schlegel, R. E., French, J., & Shehab, R. L. (1995, May). *Astronaut performance during preflight, in-orbit, and recovery.* Paper presented at the Aerospace Medical Association 66th Annual Scientific Meeting, Anaheim, CA.
Schlegel, R. E., Shehab, R. L., Gilliland, K., Eddy, D. R., & Schiflett, S. G. (1995, May). *Astronaut baseline practice schedules for the NASA performance assessment workstation (PAWS).* Paper presented at the Aerospace Medical Association 66th Annual Scientific Meeting, Anaheim, CA.
Starmer, G. (1985). Antihistamines and highway safety. *Accident Analysis and Prevention, 17,* 311–317.
Steyvers, F. J. J. M. (1987). The influence of sleep deprivation and knowledge of results on perceptual encoding. *Acta Psychologica, 6,* 173–187.
Vardeman, S. B., & Jobe, J. M. (1999). *Statistical quality assurance methods for engineers.* New York: Wiley.
Wheeler, D. J., & Chambers, D. S. (1992). *Understanding statistical process control.* Knoxville, TN: SPC.

Randa L. Shehab is assistant professor of industrial engineering at the University of Oklahoma, Norman, Oklahoma. She completed her Ph.D. in industrial engineering with an emphasis in human factors at the University of Oklahoma in 1995.

Robert E. Schlegel is professor of industrial engineering at the University of Oklahoma, Norman, Oklahoma. He received his Ph.D. in industrial engineering with an emphasis in human factors from the University of Oklahoma in 1980.

Date received: June 14, 1999
Date accepted: April 27, 2000