JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1995, **28**, 349–370 NUMBER 3 (FALL 1995)

USING STATISTICAL PROCESS CONTROL TO MAKE DATA-BASED CLINICAL DECISIONS

AL PFADT

NEW YORK STATE INSTITUTE FOR BASIC RESEARCH IN DEVELOPMENTAL DISABILITIES

AND

DONALD J. WHEELER

STATISTICAL PROCESS CONTROLS, INC.

Applied behavior analysis is based on an investigation of variability due to interrelationships among antecedents, behavior, and consequences. This permits testable hypotheses about the causes of behavior as well as for the course of treatment to be evaluated empirically. Such information provides corrective feedback for making data-based clinical decisions. This paper considers how a different approach to the analysis of variability based on the writings of Walter Shewart and W. Edwards Deming in the area of industrial quality control helps to achieve similar objectives. Statistical process control (SPC) was developed to implement a process of continual product improvement while achieving compliance with production standards and other requirements for promoting customer satisfaction. SPC involves the use of simple statistical tools, such as histograms and control charts, as well as problem-solving techniques, such as flow charts, cause-and-effect diagrams, and Pareto charts, to implement Deming's management philosophy. These data-analytic procedures can be incorporated into a human service organization to help to achieve its stated objectives in a manner that leads to continuous improvement in the functioning of the clients who are its customers. Examples are provided to illustrate how SPC procedures can be used to analyze behavioral data. Issues related to the application of these tools for making data-based clinical decisions and for creating an organizational climate that promotes their routine use in applied settings are also considered.

DESCRIPTORS: statistical process control, data analysis, methodology, applied behavior analysis, developmental disabilities

Statistical process control (SPC) includes a number of simple statistical procedures and problem-solving techniques with powerful applications in industrial manufacturing operations. In these contexts, SPC is used to detect patterns of variation in the production process that must be corrected in order to ensure that quality control standards are achieved and to implement a plan leading to continual product improvement. The introductory text by Wheeler and Chambers (1992) describes seven basic tools for accomplishing these objectives, including (a) the construction of histograms (e.g., bar charts or stem-and-leaf plots), (b) running records that provide a graphic representation of the production process, (c) use of control charts (described below), and (d) procedures for setting the process aim that provide for a more in-depth analysis of the quantitative information used to monitor productivity and to evaluate the quality of the product produced. In addition, three organizational tools for problem solving—flowcharts, Pareto charts, and cause-

Preparation of this article was supported in part by funds from the New York State Office of Mental Retardation and Developmental Disabilities. We are grateful to Paul Patti and John Niederbuhl for providing the behavioral data used to illustrate the clinical applications of SPC methods and to Ira Cohen, Vicky Mariocca, Valerie Mazza, Gene Sersen, and Vicki Sudhalter for assistance in preparing the manuscript for publication. We dedicate this article to the memory of W. Edwards Deming, whose system of profound knowledge provides the basis for successful applications of statistical process control.

Requests for reprints should be addressed to Al Pfadt, Clinical Administrator of the Intensive Behavioral Treatment Program, New York State Institute for Basic Research in Developmental Disabilities, 1050 Forest Hill Road, Staten Island, New York 10314.

and-effect diagrams—are useful in tracking down the sources of variability that result in special causes of variation. These are considered to be exceptions to the manner in which the manufacturing system normally functions. A fundamental objective in using SPC tools is to distinguish patterns of variation due to these special (assignable) causes of uncontrolled variation from patterns of controlled variation due to common (systemic) causes, which are regarded as inherent features of how the manufacturing process is organized and operated.

Intervening in a work process that is "in statistical control" and therefore displays only common causes of variation is tampering (illustrated by the funnel experiment described in Scherkenbach, 1991, pp. 41–55) and will result only in greater variability. This represents a "false alarm" and is similar to the Type I error made in conventional hypothesis-testing research. Furthermore, such a faulty analysis often leads to a tendency to blame workers for the production of defects that are due to common causes of variation. These can be corrected only by the actions of management in changing how the system fundamentally operates. This point is illustrated by Deming's infamous red bead experiment, described by Geller (1992) in his introduction to a special issue of the *Journal of Applied Behavior Analysis* (Fall 1992) devoted to applications of performance management in business and industry. This was also noted by Mawhinney (1992) in his discussion of the implications of Deming's management philosophy for organizational behavior management. Type I errors can lead to increased frustration on the part of supervisors as well as workers, who both are trying to fix a system that is operating normally. On the other hand, failure to correctly identify special causes of variation when they are present and to track down and correct the extraneous influences contributing to these defects results in a costly need to undo the damage done by faulty production. These Type II errors also result in missed opportunities to avoid such defects in the future.

As described in introductory texts by Gitlow and Gitlow (1987), Kane (1989), Montgomery (1985), Pyzdek (1989), and Wheeler and Chambers (1992), SPC offers an approach for evaluating the significance of changes associated with planned interventions or with the spontaneous occurrence of uncontrolled variables. It combines the rigor and objectivity of a statistical analysis of data with the sensitivity of clinical judgment developed by the behavior-analytic tradition that favors visual inspection of characteristics in the time series of scores from an individual subject (see Baer & Parsonson, 1981). This is accomplished by constructing a control chart, in which the average values obtained during a trial period are used to compute a measure of *location* or central tendency (the central line) and control limits, based on a measure of *dispersion* or variability in the scores obtained. The running record serves as a graphic representation of the work process. Use of control charts provides objective criteria (i.e., scaling factors) for "eyeballing" the data in much the same manner that other judgmental aids described by Bailey (1984) and Birkimer and Brown (1979) assist in the visual interpretation of quantitative information. Accordingly, use of these procedures allows the investigator to avoid choosing between either a statistical analysis or a visual inspection of time series data. Using control charts, it is possible to do both types of analysis without making the questionable assumptions that statistical analyses sometimes require (i.e., assuming that there is no autocorrelation among scores or that scores are normally distributed, etc.; see Pfadt, Cohen, Sudhalter, Romancyzk, & Wheeler, 1992, and Wheeler, 1990, for a fuller consideration of these issues).

The Logic of a Control Chart Analysis

The conceptual foundation for the use of control charts was established by Walter Shewart during his pioneering research at Bell Laboratories (see Shewart, 1939/1986) and is described in detail by Wheeler and Chambers

(1992). The upper and lower control limits of the output of a work process are established by converting measurements of the dispersion of scores in samples of some quality characteristic of the manufactured items into sigma units. Sigma units reflect the distribution of scores that can be expected to lie on either side of the central line (the mean of the samples), in terms of standard deviation units. Control limits are used to establish the parameters of common causes of variation. These limits specify a range of values to be expected if only those factors intrinsic to the normal operation of the production process affect the results obtained. A prediction is made about how the process should continue to function if only common causes of variation are operative. The control limits are extended and the subsequent output of the work process is monitored by applying certain decision rules (described below). These objective guidelines are used to determine if and when the assumption of chance variation has been violated and the effects of special causes of variation are apparent.

Some of the decision rules commonly used in an SPC analysis of time series data are described in Wheeler and Chambers (1992) and have been reproduced by Mainstone and Levy (1987, p. 15). These criteria are based on the applicability of probability theory for evaluating the likelihood of obtaining certain patterns of scores by chance, using the control limits and central line of the control chart as scaling factors. For example, a special cause of variation is suggested whenever eight successive values fall on the same side of the central line or whenever two of three points are on the same side of, and more than two standard deviation units away from, the central line.

The use of sigma units provides a basis for shifting from arbitrary measurement units that describe the characteristics of a representative sample to standardized units that provide estimates of the relevant characteristics of a process parameter. It is important to emphasize, however, that the use of three-sigma limits as the foundation for taking action on the signals detected by control charts is not based solely on probability theory, although it is supported by the application of mathematical principles derived from sampling theory. Shewart (1939/1986) stated that three-sigma limits were chosen because they provided practical guidelines for determining when it was worthwhile to invest the resources of an organization in looking for assignable causes of the abnormal patterns of variation that are detected by their use. Wheeler and Chambers (1992) observed that "the strongest justification of three-sigma limits is the empirical evidence that these three-sigma limits work well in practice—that they provide effective action limits when applied to real world data" (p. 60).

Differences between contingency-shaped and rule-based guidelines for making data-based decisions were described by Killeen (1978), who considered the relative advantages and disadvantages of each. The objective criteria that guide an SPC analysis of data sets function as rules that instruct the investigator to attend to particular aspects of the data. We provide two clinical examples below to illustrate this process. In this way, characteristics of a subject's data come to serve as discriminative stimuli for taking corrective actions on those environmental influences that are associated with problematic behaviors. If these interventions are successful, their use will be reinforced. If they are not successful, their use will be extinguished (and deservedly so). Therefore, SPC guidelines allow for the change agent's repertoire of problem-solving skills to come under the control of the data generated during planned interventions (see Saunders & Saunders, 1994). In this regard, the tools of SPC also lead to contingency-shaped decision rules and problem-solving strategies, as will be illustrated below.

The Organizational Context for SPC Applications

Taking action on the information provided by control charts and other SPC problem-solving tools requires an organizational culture that is committed to making data-based decisions. This necessitates that technical and management systems become interdependent and well integrated in order for meaningful changes to occur (see Rubinstein, 1984). Implementing the principles of Deming's management philosophy (Deming, 1986, 1993) is one way to achieve this interdependency. However, there are points of disagreement as well as areas of considerable overlap between Deming's "system of profound knowledge" and the metatheoretical assumptions that underlie the applied analysis of behavior. Contributors to special issues of the *Journal of Organizational Behavior Management* (Vol. 9, No. 1) and the *Journal of Applied Behavior Analysis* (Vol. 25, No. 3) expressed a range of opinions about the compatibility of these two approaches. It is beyond the scope of this article to consider this issue in detail. Nevertheless, the application of SPC will have only a limited impact unless it is supported by the top levels of management within an organization. This is also true for any of the technological innovations developed by applied behavior analysis. We will consider issues related to incorporating SPC into routine clinical practice in the last section of this article.

Brache and Rummler (1988) described three levels of quality improvement in manufacturing settings related to individual workers, production processes, and organizational variables. According to Brache and Rummler, "the organization, process, and individual levels are interdependent, linked together in a total quality system that ultimately determines the quality of an organization's products and services" (p. 46). In this article we will be primarily concerned with describing how control charts can be used to evaluate behavioral data when the focus is on changing the performance of an individual client. However, SPC tools can be introduced in an applied setting in order to change the behavior of a clinical team, so that members use a more data-based approach to treatment planning and decision making. Finally, it is possible to regard Deming's management philosophy as a set of rules for consultants to follow in order to promote systems-wide changes within an entire organization, as discussed by Saunders and Saunders (1994). The data-analytic and problem-solving tools of SPC described below can be used to implement a process of continual improvement based on the plan-do-study-act cycle detailed by Scherkenbach (1991, pp. 63–81). Issues related to incorporating SPC into routine clinical practice in applied settings will be discussed in the final section of this article.

CLINICAL APPLICATIONS OF SPC METHODS

The use of SPC in clinical contexts is based on the premise that variability in the occurrence of target behaviors can be analyzed by using procedures that have been developed in industrial settings as rules of thumb for evaluating time-ordered measurements in the output of a work process. Redmon (1992) observed that "changes in steady state depicted on an SPC control chart represent changes in controlling variables in much the same way as variations in graphic data patterns signal changes in performance as a function of environmental phenomena in applied behavior analysis" (p. 547). A similar observation was made over 40 years ago by Wilson (1952/1990), who reported that attempting to control all relevant sources of variation in an experimental situation "does not necessarily eliminate all variability in the results because there will be a very large number of variables left which individually have small effects but which collectively produce a scatter in the results" (p. 262). Noting that this scatter could be interpreted as essentially random (or in statistical control), Wilson then provided a number of examples (see pp. 263–272) to illustrate how the statistical principles developed by Shewart (1939/1986) for industrial quality control could be applied to analyze the results of planned interventions. Pfadt et al. (1992) and Pfadt and Wheeler (1993) have shown how these statistically based decision criteria can be used to carry out the interpretive guidelines discussed by Parsonson and Baer (1986) in evaluating changes in level and trend in behavioral data during baseline or during treatment in clinical settings.

The direct application of SPC procedures is most straightforward in those cases in which some aspect of a client's behavior can be considered to be defective and the goal is to identify and correct any ecobehavioral variables that are associated with these defects. However, such an eliminative approach (Hawkins, 1986) is essentially reactive and involves "scraping burned toast," because the targeted behavior must occur often enough for these corrective strategies to be implemented. For this reason, Hawkins recommended constructional approaches to treatment (originally described by Goldiamond, 1974) that attempt to expand the repertoire of the person displaying the problem behavior by teaching new skills. From this perspective, the change agent will be interested in the ability of control charts to detect *when* a new level of performance (or training criteria) is achieved, so that the success of this enterprise can be monitored in real time. Changes in treatment can then be based on this information. For example, the training program might be modified if specified outcomes are not achieved by a target date. This application is similar to the use of standard celeration charts in precision teaching (Lindsley, 1992). Application of control charts in the context of a constructional approach to treatment is also an effective strategy for operationalizing a changing criterion design (Tawney & Gast, 1984), whereby the upper control limit computed during the client's preintervention baseline is used to determine the changed training criterion in effect during the treatment phase.

Control charts can be applied in three different but interrelated areas of treatment planning and clinical decision making: (a) The decision rules described by Wheeler and Chambers (1992) for detecting changes in level, trend, or both can be applied to determine whether the criteria for baseline stability have been achieved or if baseline observations indicate the presence of special causes of variation; (b) control charts can be used retrospectively, as judgmental aids, to provide objective guidelines for determining whether or not an intervention was effective in comparison with the preintervention baseline data; and (c) control charts can be used to monitor the effects of interventions in real time and to implement a changing criterion design that selects the values used to indicate a change in clinical status on the basis of pretreatment functioning and not in terms of some arbitrarily imposed criteria. Here we will be concerned with issues related to selecting, constructing, and interpreting control charts to evaluate baseline stability. Elsewhere (Pfadt et al., 1992; Pfadt & Wheeler, 1993), we have illustrated the second and third applications. We will also consider how SPC problem-solving techniques (Pareto charts, cause-and-effect diagrams, and flowcharts) can be used in clinical settings to identify the source of any special (abnormal) patterns of variation detected in the running record. First, however, we will consider some measurement issues regarding the behavioral stream as similar to a work process.

The Behavioral Stream As a Work Process

Mawhinney (1992) and Redmon (1992) showed how, when viewed from a systems perspective, both SPC and applied behavior analysis are alike in their attempts to analyze the causal systems affecting the measurable properties of the "process" under investigation into those components responsible for transforming inputs into outputs. Welch (1992) expressed a reservation to the application of SPC for behavioral data by noting that "after all, people are not widgets and their errors cannot be treated as scrap" (p. 10). However, as noted previously, variability in the measures of both behavioral and work processes can be analyzed by means of control charts. The ecobehavioral model of Morris and Midgley (1990) described below also calls attention to similarities between the behavioral stream and a work process. This reflects the historical influences of ecological psychology and field theories of behavior that gave rise to this perspective, as Morris and Midgley discuss in their article.

The products of a work process are comprised of relatively static properties that can be measured as fixed characteristics to yield variable or attribute data. In the latter case, "measurement consists of evaluating whether an individual part or item has a particular feature, i.e., is a part defective or nondefective" (Kane, 1989, p. 26), although the number of blemishes (or defective features) on a particular part may also be counted. However, the behavioral stream is a continuous process. Although it is possible to speak of "behavioral units" (Thompson & Zeiler, 1986), these are not static products (in this respect, people really aren't widgets!) and these units cannot be considered as discrete items. In behavior analysis the response class (Skinner, 1935), rather than a specific instance of behavior, is the fundamental unit of analysis. Furthermore, it is assumed that functional relationships rather than topographical features are the most important characteristics in describing the "natural lines of fracture" (Skinner, 1935) that determine the flow of the behavioral stream. Barrett, Johnston, and Pennypacker (1986) observed that "response classes defined by empirical verification of functional homogeneity across their various topographies constitute the maximally efficient pretreatment assessment baselines from which to select and against which to evaluate the effects of the remediation procedures" (p. 170). Units of measurement that can be used as appropriate standards for quantifying clinically relevant dimensions of behavior have been described by Johnston and Pennypacker (1993, chap. 5) and will only be listed here. They include frequency, countability (rate), celeration (the second derivative of frequency), duration, latency, and interresponse time (IRT).

From an SPC perspective, frequencies of target behaviors (even though a well-defined observation code might be employed to yield reliable and accurate data) are regarded as attribute measures. Three inherent frustrations associated with attribute data have been described by Wheeler and Chambers (1992): (a) Control charts constructed from small counts in which the average per subgroup is less than one are insensitive, necessitating long measurement intervals to detect enough nonconformities to establish representative values; (b) such binary data have an all-or-nothing feature that obscures information about the underlying dimension; and (c) there is ambiguity about whether the data-collection process or the behavior of the production process is responsible for the different measures obtained. Although Johnston and Pennypacker (1993) acknowledge that, for some purposes, frequency of all target behaviors identified in a sampling interval (or the derived measure, rate) may not provide the most clinically relevant information (see also Baer, 1986), they do not ascribe a lesser status to count data than to other measures of behavioral dimensions (e.g., duration or IRT). However, Johnston and Pennypacker do give primacy to mechanical transducers relative to human observers for many purposes and they recommend that "given their strategic advantages, researchers should fully consider the possibilities for using machines as transducers before turning to humans" (p. 118). A similar conclusion was reached by Pfadt and Tryon (1983), who described a variety of uses for mechanical transducers in clinical settings. In this sense, then, behavioral assessment can provide the equivalent of "measurement by gauges" that results in variable data. It is also possible to use measures of duration (as well as latency and IRT) to provide variable data that can be expressed in the dimensions of natural science (see Barrett et al., 1986, Table 1, p. 174, for the properties and units involved). We will illustrate certain advantages to using variable data (such as IRTs) instead of frequency counts of operationally defined target behaviors (attribute data) in the examples presented below.

Analyzing Baseline Stability

Stable rates of responding during baseline are important for the basic (experimental) as well as the applied analysis of behavior for theoretical and practical reasons. Conceptually, stability has traditionally been regarded as an index of the extent to which an investigator has established control over relevant variables that affect responding in the experimental condition (Sidman, 1960). Pragmatically, stability during baseline is necessary if a subject's preintervention time series is to be useful in detecting subsequent changes in level, trend, or both during treatment. Because the effectiveness of treatment within the behavior-analytic tradition is typically determined by visual inspection of graphed data for an individual subject, stability during baseline enhances the investigator's ability to detect treatment effects. Baer and Parsonson (1981) and Kazdin (1982), among others, have discussed the interpretive problems posed by variability during baseline. Hartmann et al. (1980) observed that "the eye has trouble distinguishing real behavior change from random fluctuations when scores are highly variable" (p. 544) and have recommended that such measures be aggregated to produce a more stable time series if it is not possible to rerun the experiment under more tightly controlled conditions. This advice, however, begs the question of determining when a baseline is stable enough to provide a useful frame of reference for evaluating subsequent effects.

The stability criteria proposed by Cumming and Schoenfeld (1960) and Killeen (1978) represent attempts to use statistical principles to answer the question: When is a baseline stable enough to proceed with the next phase of treatment? Cumming and Schoenfeld considered responding to be stable enough if the difference between the means of two consecutive 3-day periods was within ±5% of the overall 6-day mean. Killeen reanalyzed the data presented by Cumming and Schoenfeld and suggested several additional criteria, including a coefficient of variation (CV = the standard deviation divided by the mean times 100) that did not exceed 14%. Elsewhere (Pfadt, Wheeler, Sersen, & Moreno, 1993), we showed how control limits can be used to inspect the classic data set presented by Cumming and Schoenfeld to provide an operational definition for determining when stability has been achieved during baseline. A baseline can be said to be stable when there is no evidence that assignable causes of variation are present for the individual values of the time series in the X chart or in the mR chart for the moving ranges, using the set of four detection rules presented in Wheeler and Chambers (1992, p. 96):

1. A single point that falls outside the three-sigma control limits.
2. Two of three points that fall on the same side of, and more than two-sigma limits away from, the central line.
3. At least four out of five successive points fall on the same side of, and more than one sigma unit away from, the central line.
4. At least eight consecutive values or 12 of 14 successive points fall on the same side of the central line.

We will apply these criteria to analyze the baselines presented below. These examples also provide opportunities to consider some measurement issues related to constuctional versus eliminative approaches to treatment. We will also illustrate the different properties of attribute versus variable data when they are analyzed by control charts.

An eliminative approach to treatment using attribute data. The data presented in Figure 1 represent daily frequency counts of operationally defined instances of "aggressive outbursts" exhibited by a 30-year-old man who functions in the severe range of mental retardation. Daily frequencies of these target behaviors (instances of hitting or kicking others, or using objects as weapons to attack other clients or staff) were obtained from incident reports completed by

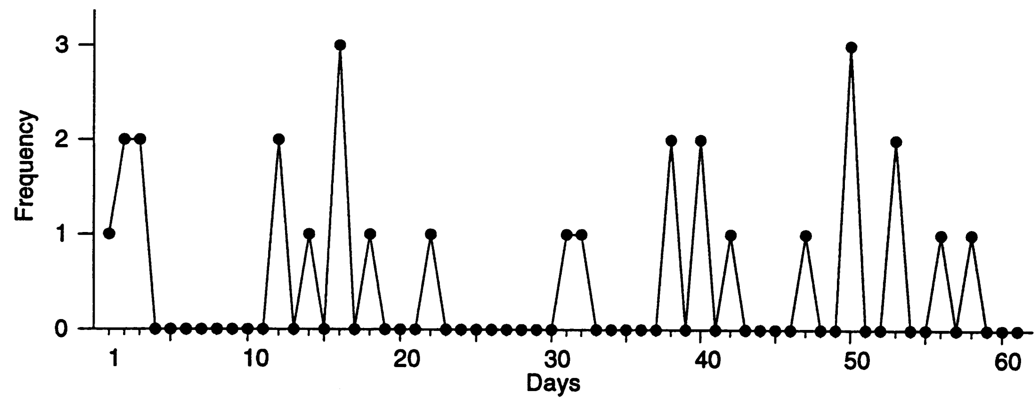

Figure 1. Daily frequencies of aggressive outbursts for Mr. X during baseline.

staff in a public residential facility as part of an approved behavioral treatment plan. They were compiled by the agency psychologist as part of a consultation at the diagnostic and research clinic where the senior author is employed.

When the graph of these data is interpreted without benefit of the judgmental aids provided by control charts, it appears that no clear "signals" (i.e., changes in level or trend) are present in this "noisy" baseline. This impression is borne out by the value of the linear regression (−0.06) computed for this time series, using procedures described by Baer and Parsonson (1981). However, when these baseline values are displayed on a control chart (see the upper panel in Figure 2), six clear signals can be detected using the four decision rules presented earlier. The central line (CL = 0.46) is the mean of the baseline time series. The upper control limit (UCL = 2.54) was calculated by first computing the mean of the moving ranges for the baseline values (i.e., the absolute values of differences between successive daily frequencies). This value (0.78) was multiplied by a conversion constant (2.66), which was then added to the value for the central line: $(2.66 \times 0.78) + 0.46 = 2.54$. The interested reader can compute the moving ranges for all of the values shown in the lower panel in Figure 2 to verify these computations (which are accurate to within rounding error) and to practice constructing control charts for similar data. For example, the first moving range is one (the difference between the daily frequency of one on Day 1 and two on Day 2) and the second moving range is zero (the frequencies on Day 2 and Day 3 are the same). Notice that the number of moving ranges will be one less than the number of data points. Some statisticians might prefer to use a C chart (see Wheeler & Chambers, 1992, pp. 271–274) to analyze data such as these. The results are essentially the same, and the individual (X) and moving range (mR) charts do not require any assumptions about the underlying distribution of scores.

The value of the conversion constant used for the X control chart was taken from the table presented in Wheeler and Chambers (1992, p. 393) for selecting bias correction factors when the sample size is two. This is the case when the moving range is used to estimate the within-sample variation if only single measurements (such as daily frequencies) are obtained (see Wheeler and Chambers, 1992, p. 48). The three-sigma UCL helped to identify two signals present in the upper panel in Figure 2. These occurred on Day 16 and Day 50 when the daily frequency of three exceeded the UCL of 2.54. An estimate of sigma in a time series of individual scores was obtained by multiplying the mean of the moving ranges by the conversion constant and dividing by three: $(0.78 \times 2.66)/3 = 0.69$. This sigma value was used to carry out the other tests mentioned earlier. A two-

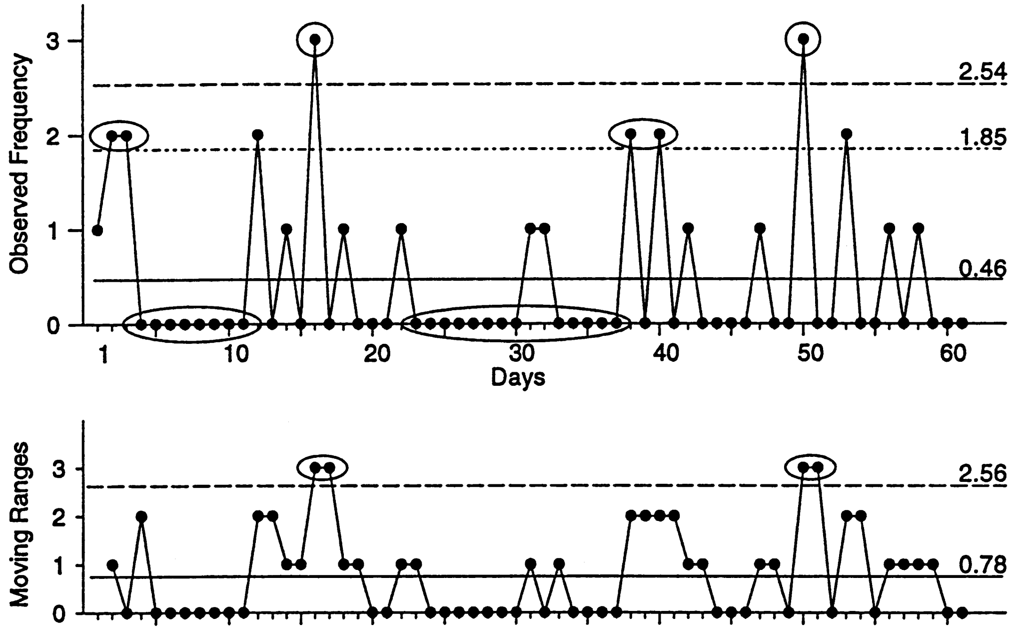

Figure 2. Control charts for the individual values (upper panel) and the moving ranges of the daily frequencies shown in Figure 1. The central line (CL = 0.46) is the mean of the time series and is represented by the solid line in the upper panel. The circled values below the CL in this panel are runs of scores that are out of statistical control, according to criteria defined in the text. The upper control limit (UCL = 2.54) is represented by the broken line in the top of this panel and is used to identify values on Day 15 and Day 50 as potential signals. The dotted line between the CL and the UCL is the two-sigma line, used to identify clusters of values (two of three consecutive values) that are potential signals. Values shown in the lower panel are moving ranges, the absolute values of differences between successive daily frequencies. The UCL (2.56) for the moving range chart is represented by the broken line above the CL (0.78). The four values in the lower panel that exceed this line identify discontinuities in the orginal time series.

sigma line was calculated by adding the value of two times sigma to the central line: $(2 \times 0.69) + 0.46 = 1.84$. Drawing this two-sigma line on the graph helped to identify two more signals in this baseline—a bad period that included Days 2 and 3 and another bad period from Days 38 to 40, where two of three consecutive values were above the two-sigma line. The central line itself was used as the frame of reference for identifying the last two signals present in this baseline. The first is indicated by a run of eight consecutive values below the central line from Day 4 through Day 11, and the second corresponds to a run of 12 of 14 values below the central line from Day 23 through Day 36.

Instead of regarding all of the variability in the baseline time series for Mr. X as "noise," use of control charts permits the identification of six signals that are presumably due to special causes of variation. In fact, the use of control charts to analyze the data presented in the upper panel of Figure 2 suggests that it is not the behavioral output of a stable process. Rather, these data have apparently been produced by at least two different processes, producing signals on the high side or the low side of the central line. The signals identified as circled values in this panel tell an investigator where to look in order to determine which ecobehavioral influences cause the observed behavior. In this sense, the control chart functions as a road map, calling attention to patterns in the data. Without these objective criteria, one risks the error of interpreting any change as if it were a signal (a Type I error that leads to tampering, as discussed earlier). This raises the question: How do we know that patterns are really signals due to special causes of variation and not simply the results of chance fluctuations? Several types of evidence support the former interpretation. The first is related to the "logic of control charts" described earlier. If one flipped an unbiased coin eight times in a row, the probability of getting eight successive heads or tails would be p = .008 ($1/2^7$ after throwing the first head or tail). Of course this outcome could have happened by chance, but a more likely explanation would be that the coin was biased. Likewise, the odds that a run of eight consecutive values above or below the central line could have happened by chance are very small, approximately one chance out of 128. However, it is more reasonable to assume that a special cause of variation altered the underlying cause system and thereby influenced the subject's performance during that time period. A simple test for determining whether a set of scores contains sequences that are not randomly distributed was presented by Wilson (1952/1990). The central line in the upper panel of Figure 2 is used as a reference point to count the clusters of scores below the mean of the distribution (0.46). For 60 observations, any number of runs below the critical value of 22 (taken from Wilson, 1952/1990, Table 9.9, p. 267) suggests a nonrandom distribution of scores. There are only 15 clusters of scores below the mean of this series (i.e., the run of eight scores from Day 4 through Day 11 is counted as one cluster). This is sufficient to reject the null hypothesis at p = .025.

The mR chart shown in the lower panel of Figure 2 also provides supportive evidence that there are meaningful signals in the running record for Mr. X presented in Figure 1. The control chart shows that there are two abrupt transitions which correspond to the moving ranges before and after the bad days identified in the upper panel of Figure 2 (Day 16 and Day 50). The UCL for the mR chart was computed by multiplying the mean of the moving ranges (0.78) by a bias correction factor (3.268) that is used to construct mR control charts (see Wheeler & Chambers, 1992, p. 48). Decision Rules 2, 3, and 4 described above are not used to analyze runs of scores on an mR chart because of the manner in which moving ranges are computed. Except for the first and the last score, each measurement contributes to two consecutive moving ranges. This would make it twice as likely to make a Type I error if tests were used to identify sequences of scores that are out of control. Additional supportive evidence could be provided by any collateral observations (reports from staff, performance on tasks, incident reports such as documentation of the need for medical treatment) that were part of the clinical records; these might suggest that bad days or periods really were different from good days or periods. Ultimately, the ability of the investigator to interpret the information generated by control charts in the context of knowledge about the process that generated the data and to use this new information to improve subsequent performance determines whether or not these out-of-control values should be regarded as meaningful signals or extraneous noise.

It is worth noting that both the X and the mR charts shown in Figure 2 call attention to bad days, when frequencies above the upper control limit were recorded. This tends to negatively reinforce the investigator's development of eliminative approaches to treatment and reduces the likelihood that these aversive stimuli will appear again in subsequent graphs of a client's performance. However, complete suppression of problematic target behaviors may not be a realistic outcome if the individual does not have an alternative behavior in his or her repertoire that can provide equivalent access to reinforcing consequences. Furthermore, graphs of frequency data may fail to identify potential signals in a time series that indicate improvements in functioning related to desirable conditions that should be enhanced or programmed to occur more predictably. We will illustrate below how a constructional approach to treatment can be facilitated by the use of IRT data.

A constructional approach to treatment. Although the use of control charts to analyze the frequency data in Figure 2 helps to differentiate signals from noise in the baseline time series, there is a serious limitation to this approach if we want to use these control limits to monitor treatment outcomes. Because incidents occur only every other day (CL = 0.46) on the average, it is not possible to compute a meaningful lower control limit or even one- and two-sigma warning limits *below* the central line. In order for improvements to be apparent, one must eliminate the target behavior for prolonged intervals (8 consecutive days or 12 of 14 successive days). More subtle treatment effects cannot be detected using these decision rules. Parsonson and Baer (1986) suggested that this is not really a problem, because only robust treatments will survive this type of selection process. However, many applied interventions may be worthwhile if they lead to more modest outcomes, which can then be made part of a more comprehensive treatment package (see Fawcett, 1991). Use of IRTs (length of time between the occurrences of target behaviors) provides a means of obtaining variable data that can be used to analyze Mr. X's baseline time series.

Each of the events shown in the upper panel of Figure 2 was recoded to provide an IRT measure using the following procedure. Because only the number of events per day and not the exact time when incidents occurred was available, the IRT was estimated by assuming that if only one event was reported, it occurred at the end of the day. Two or three events per day were assumed to be equally spaced throughout the day. For example, three events on the same day were coded as IRTs of 0.33, 0.33, and 0.33. Of course, if events had been coded in real time it would not have been necessary to make these simplifying assumptions and a source of potential bias or confusion would have been eliminated. Furthermore, in the absence of real-time measurements of the duration of episodes when the target behaviors were displayed, it was not possible to measure each IRT precisely or to calculate the total IRT interval when responding could have occurred. A precise measure of rate (count divided by total IRT) therefore could not be computed, as recommended by Johnston and Pennypacker (1993). Nevertheless, the IRT data presented in Figure 3 permit a consideration of how a constructional approach to treatment (see Goldiamond, 1974; Hawkins, 1986) differs from an eliminative approach. The three-sigma upper control limit (UCL = 8.20) shown in the upper panel of Figure 4 was computed in the manner described previously. The value of the CL, 2.04, was added to the product of the mean of the moving ranges (2.31) shown in the upper panel of Figure 4 and the bias correction factor (2.66) used for X charts: 2.04 + 2.31 (2.66) = 8.20. Two of the IRT values shown in Figure 4 exceeded this three-sigma criterion.

There is no time axis on an IRT chart, although each event corresponds to a particular interval of time. However, by comparing the IRT values shown in Figure 3 to the frequency data presented in Figure 2, the time periods corresponding to the out-of-control signals shown in Figure 3 can be identified. The upper control limit provides an empirically derived, objective criterion for determining when the run of good days on an IRT chart is sufficiently different from other patterns seen in a running record to qualify as an out-of-control signal. In this case, more than 8 consecutive days without an incident are required. This happened on two occasions, once from Day 4 through Day 11 and again from Day 23 through Day 30, as shown in Figure 3. Moreover, use of IRT as an outcome measure calls attention to treatment strategies that enhance the person's ability to function in a socially desirable manner (e.g., by teaching self-control strategies that lessen the person's vulnerability to triggering events that might lead to the expression of problematic behaviors, as discussed by Gardner & Cole, 1989). Whereas there was little room for improvement in the control chart for daily fre-

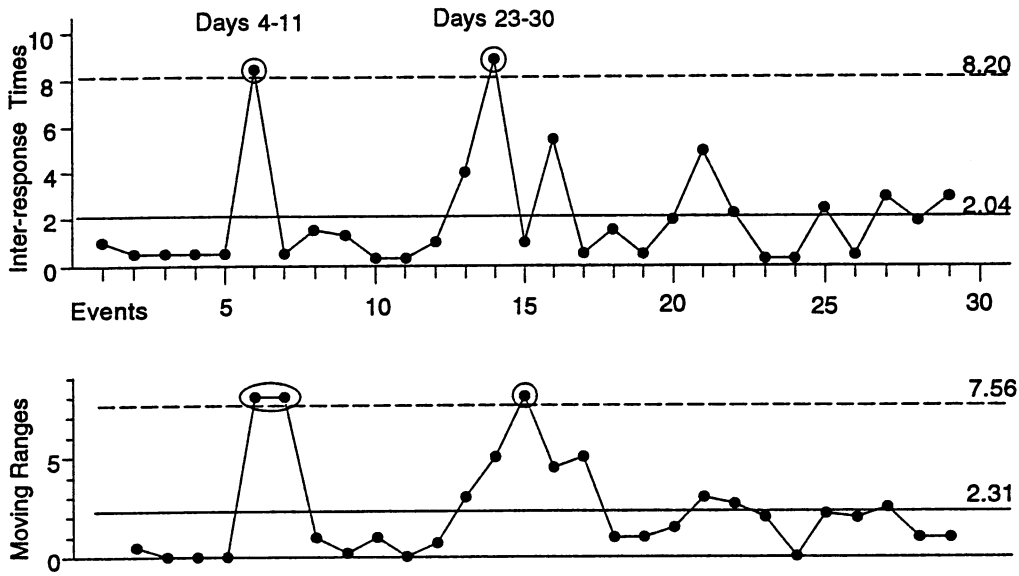

Figure 3. Control charts for the IRTs calculated for the daily frequencies shown in Figure 2, according to procedures described in the text. The circled values in the upper panel correspond to prolonged intervals from Days 4 to 11 and Days 23 to 30 when no aggressive outbursts were reported. The moving range chart in the lower panel shows three circled values that identify breaks in the original time series.

quencies shown in Figure 2, there is unlimited room for improvement in an IRT chart for the same data. It is also possible to construct meaningful one- and two-sigma warning limits for these IRT scores that can then be used to evaluate changes in functioning. For example, two of three consecutive IRTs more than 6 days apart indicate than an improvement has taken place, according to Decision Rule 2 presented above. Likewise, four of five consecutive IRTs greater than the one-sigma value (four) signals a change in functioning. In clinical terms, this indicates that the person was able to "regroup" more quickly after losing inhibitory control over a problematic behavior and was then able to display another run of good days. This strat-

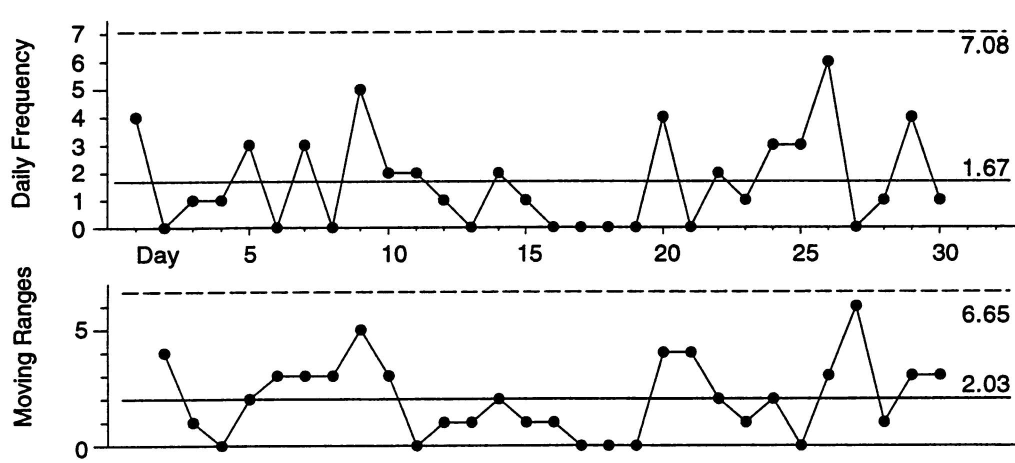

Figure 4. Control charts for the number of times per day that restraints were used with Ms. Y, as shown in the Appendix. Both the chart for individual values (upper panel) and the moving range chart (lower panel) are in statistical control.

egy was used successfully to evaluate drug treatment programs by Mann, Charuvastra, and Murthy (1984), who calculated the length of time between relapses as an outcome measure. It seems to offer promise in other areas as well (see Johnson & Pennypacker, 1993, chap. 5, for a discussion of IRTs per opportunity as a dimensional property of behavior).

The moving range chart of the IRT data for Mr. X is presented in the lower panel of Figure 3. The out-of-control signals in this chart reflect abrupt transitions from bad to good periods and support the interpretation offered above, that the periods from Days 4 through 11 and Days 23 through 30 are meaningful signals.

Another Application

The principles discussed above can be applied to analyze the data presented in the Appendix, which were used to evaluate the treatment of severe pica displayed by Ms. Y, a 33-year-old woman with severe mental retardation who resided in a developmental center. The use of physical restraints (a protective helmet and special "posey" mittens to restrict hand movements) was authorized as part of a behavioral treatment plan that attempted to prevent the ingestion of dangerous materials. This was accomplished by assigning a staff person to provide constant surveillance, thereby protecting Ms. Y's safety. On two occasions in the year prior to authorization of this plan, surgery was necessary to remove large metal objects (as well as a pair of surgical gloves) from her stomach; Ms. Y had obtained these items by persistently searching her environment for objects to swallow. Protective equipment was also used to control other self-destructive behaviors (repeated attempts to pull off and swallow her fingernails, self-injurious head banging, hitting herself with her fists). These incidents occurred episodically in spite of behavioral and pharmacological interventions that were carried out in a specialized treatment unit where she lived and attended day programming. As required by state law, an incident report was filled out each time restraints were applied. This report documented the length of time restraints were used and described the circumstances that led to their application. The second and third columns in the Appendix reflect the data reported by each work shift on the total duration and the number of times that restraints were used each day.

Instead of analyzing specific target behaviors (e.g., instances of pica or attempts to ingest foreign objects, head banging, nail pulling, etc.), the use of protective equipment by staff was selected as an outcome measure during a period of baseline assessment in April. This was based on the judgment of staff members who were familiar with Ms. Y that different behavioral topographies reflected similar management problems. Accordingly, an operational definition of a *severe behavioral incident* was provided by the use of restraints to control a problematic situation. This measurement strategy reflects an emphasis on the system rather than the individual, given that factors other than Ms. Y's behavior might have contributed to the use of restraints on a particular occasion (e.g., the presence of certain objects in the setting, the judgment of staff about the dangerousness of the situation, etc.). However, eliminating the use of restraints was a top priority of Ms. Y's treatment plan. Therefore, monitoring changes in this variable reflected progress towards attaining an important clinical outcome. This approach to treatment is consistent with the behavior engineering model developed by Gilbert (1978), which distinguishes *behavior* (i.e., the movements of a person that can objectively be recorded) from *accomplishments* (defined in terms of the tangible results or products of behavior). In Gilbert's terms, this measurement strategy seeks to monitor changes in performances of the ecobehavioral system (see Morris & Midgley's, 1990, model described below) that includes the client, the staff, and the physical setting in which treatment takes place. The use of restraints is a measure of the competence of this system in coping with the demands of Ms. Y's challenging behaviors, given the resources that are available for treatment in a particular context. Changes in performance were monitored in this case by measuring relevant accomplishments in reducing the use of restraints (e.g., number and duration of applications, length of time between applications). The Appendix shows that restraints were used 50 times during the month of April, for an average duration of 45 min. No changes in medication or behavioral treatment were made during this 30-day period.

Daily frequencies for the number of times that restraints were applied during the month of April were used to construct the X and mR charts shown in Figure 4. According to the decision criteria illustrated above, Ms. Y's baseline appears to be stable. All values in the X and mR charts are in statistical control, and there don't seem to be any changes in level or trend. However, as with the frequency data presented earlier for Mr. X (see Figure 2), there are practical limitations in using these control charts to evaluate treatment outcomes. There are no meaningful sigma limits below the central line for the frequency chart shown in the lower panel of Figure 5, because CL (1.67) minus one sigma (1.80) is a negative value. Therefore, it is necessary to obtain eight consecutive values below the central line (frequencies of one or less) before improvements can be reliably detected. Furthermore, due to the day-to-day variability in this time series, the upper control limit (7.08) is quite high. This means that a daily frequency of eight or more is necessary to detect a worsening of behavior on a particular day.

The duration measures reported in the Appendix indicate that these values are highly redundant in comparison with the frequency data. This impression is borne out by the control chart for duration measures shown in the upper panel of Figure 5, which also indicates that the baseline for Ms. Y is stable (i.e., in statistical control). The moving range chart for the duration measures reported in the Appendix is also in statistical control, as shown in the lower panel of Figure 5.

The use of IRT as an outcome measure for Ms. Y is illustrated in Figure 6, a control chart of IRT scores computed for the shift values shown in the Appendix. Here the IRT score for

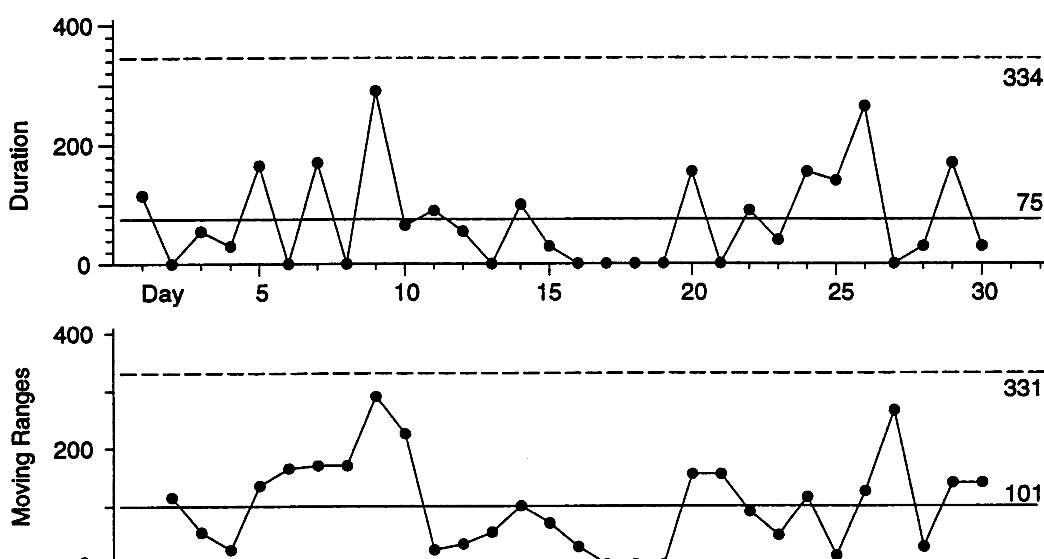

Figure 5. Control charts for the duration (in minutes) of use of restraints with Ms. Y each day, as shown in the Appendix. Both the chart for individual values (upper panel) and the moving range chart (lower panel) are in statistical control.

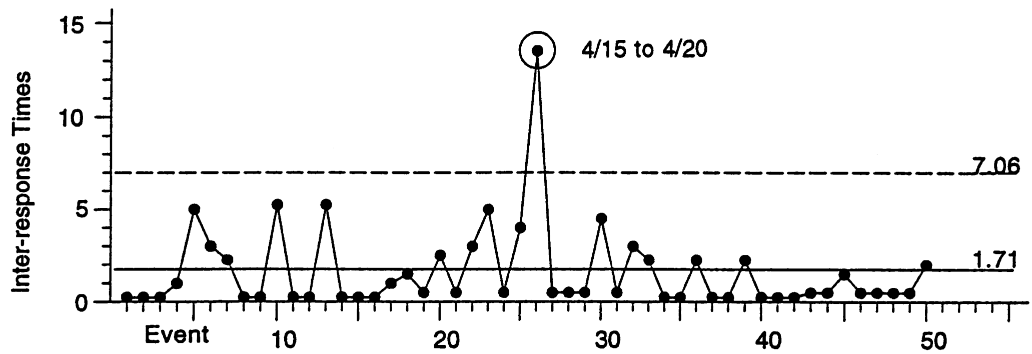

Figure 6. Control chart for the IRTs calculated for the data on the use of restraints shown in the Appendix, according to procedures described in the text. The circled value corresponds to a prolonged interval (over 13 consecutive work shifts) when restraints were not used between April 15 and April 20.

a particular incident refers to the number of consecutive work shifts between successive applications of restraints. On occasions when it was necessary to use restraints more than once per shift, it was assumed that they were equally spaced throughout the shift, because information was not available indicating when restraints were applied in real time. For example, on April 1, restraints were applied three times on the day shift for an IRT of 0.33 for each occurrence. If only one application was necessary per shift, it was assumed to occur at the end of the shift, as a simplifying assumption so that IRTs could be computed consistently. For example, a value of 1 was assigned for the occurrence on the evening shift of April 1. The next application of restraints (on the day shift of April 3) received an IRT of 5. The X control chart for these IRT values is presented in Figure 6. This chart calls attention to a prolonged interval when restraints were not used (from the end of the evening shift on April 15 through the end of the day shift on April 20). This period of improved functioning (over 13 consecutive shifts when restraints were not needed) was not apparent in the frequency chart shown in Figure 4, again illustrating an advantage of the constructional approach to treatment. The control chart for IRT values directs the clinical team to look for explanations (special, assignable causes) to account for this period of improved functioning that might be incorporated into Ms. Y's treatment plan. Parsonson and Baer (1986) recommended a similar strategy, observing that "an analysis of baseline, to see what it contains that sometimes accomplished what the researcher intends to accomplish later with a specific intervention, may show the researcher an even better intervention" (p. 171).

As noted previously, applied behavior analysts have regarded rate measures as the best indicator of response strength (see Barrett et al., 1986; Johnston & Pennypacker, 1993). For those more comfortable with this unit of analysis, IRT scores can be converted to instantaneous rates by taking the reciprocal of IRT (i.e., instantaneous rate = 1/IRT). These instantaneous rate scores are displayed on the control chart shown in Figure 7. None of the individual values are above the upper control limit of 4.92 computed for this control chart. However, during the day shift on two occasions (April 9 and April 26), three consecutive instantaneous rate measures exceeded the two-sigma warning limit of 3.87. This suggests that special influences may have been present on those days. The SPC problem-solving tools described below might be helpful to a clinical team in determining which factors might have contributed to the high rates reported on those two shifts.

SPC Problem-Solving Tools

Several procedures have been described in the SPC literature (see Gitlow & Gitlow, 1987,

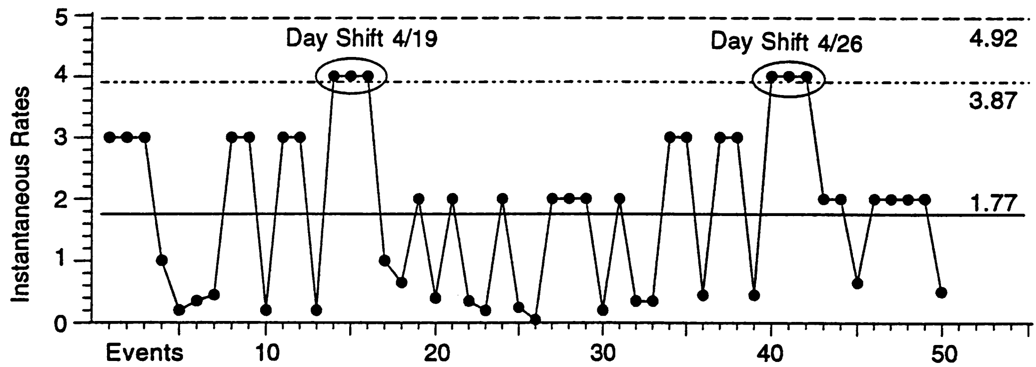

Figure 7. Control chart for instantaneous rate (1/IRT) values shown in the Appendix. Circled values exceed the two-sigma line (3.87) for the day shift on April 19 and April 26; these were used to identify clusters of values (two of three consecutive values) that represent potential signals.

chap. 5; Kane, 1989, Part II, chap. 8–13; Pyzdek, 1989, chap. 7) that might be relevant for applied behavior analysts interested in identifying the causes of erratic performance. These data-based problem-solving tools include Pareto diagrams, flowcharts, and scatter plots or histograms to identify variables that might be responsible for variation that could be better controlled in order to improve product performance or to enhance quality control. Each will be described briefly here as they might be applied in clinical settings as adjuncts to more traditional functional analytic strategies with similar purposes. More detailed descriptions of how to adapt these procedures for use in human service organizations can be found in Albin (1992) and Mawhinney (1992).

The problem-solving tools mentioned above can all be regarded as different approaches to constructing cause-and-effect diagrams, in which the basic intent is to identify all of the potentially relevant influences that contribute to a problematic situation (either a specific target behavior or a particular type of dysfunction). A generic fishbone diagram (also known as an Ishikawa diagram) may provide a useful organizing framework to focus the group's problem-solving efforts, but team members should be free to fill in the specifics for each branch on the tree themselves. For the purpose of analyzing behavior problems, the categories of an ecobehavioral model described by Morris and Midgley (1990) might be helpful in stimulating group exploration of the universe of possible influences that could contribute to maladaptive functioning. Morris and Midgley identified five types of influence (person variables, the stimulus environment, a historical context, the media of contact, and the current context) to be considered in analyzing a particular behavior. These variables represent varying degrees of complexity within the interbehavioral field. The use of histograms, scatter plots, and Pareto charts represents a more quantitative approach to the analyses of behavior problems, analyses that use frequency counts of specific components of a problem to investigate potential causes of the variability observed. These are similar to the guidelines for conducting a behavioral analysis described by Groden (1989), among others. The use of scatter plots to explore associations between variables that might not be apparent by inspecting the frequency × time plot used to construct a running record has been recommended by Touchette, MacDonald, and Langer (1985) for analyzing behavioral data. Likewise, Haring and Kennedy (1988) have recommended a graphic display of task-analytic data that permits a more fine-grained analysis of the errors that contribute to defective task performance. A less technical introduction to this topic and the other graphic problem-solving procedures discussed above is provided by Moran, Talbot, and Benson (1990). Detailed examples of how to use the tools of SPC for planning and solving problems within human service organizations to implement the principles of total quality management can be found in Albin (1992). Change agents interested in restructuring the chain of command so that an organization's philosophy is more conducive to the application of SPC for clinical decision making will find Albin's applications of Gilbert's (1978) performance engineering matrix particularly helpful (as illustrated below).

INCORPORATING SPC INTO ROUTINE CLINICAL PRACTICE

It has been our experience that SPC makes demands on both workers and management. Leaders within an organization must create a climate that empowers both of these groups to respond appropriately to the orderly accumulation of knowledge that is made possible by the systematic use of the data-analytic and problem-solving tools we have described in this article. If the metacontingencies (Glenn, 1991; Mawhinney, 1992) operating within a particular organization do not result in sufficiently reinforcing consequences whenever action is taken to correct problems uncovered by applying these tools, their use will become aversive and these procedures will gradually be eliminated. This is not because SPC is overly labor intensive. For example, a hand-drawn version of the control charts shown in Figure 2 was completed in less than 30 min, including time spent calculating control limits and inspecting the graph to identify data points that were out of statistical control. However, if adequate technical and support systems have not been established to allow workers to correct the problems identified (whenever this is possible), then this will be seen as "busy work" that is neither productive nor in the best interests of the customer (i.e., the clients served by the agency).

We have previously noted that Deming's (1986, 1993) management philosophy has been recognized as a means for creating an organizational culture that is committed to making these types of data-based decisions. Saunders and Saunders (1994) have analyzed Deming's management philosophy from a behavior-analytic perspective and have concluded that "instituting change through SPC methods (changes that later can be said to reflect adherence to Deming's 14 points) results in bringing workers and managers into more direct contact with the reinforcement contingencies operating on the manufacturing process" (p. 121). This same outcome could be achieved in health care settings if leaders were able to make SPC an integral part of the decision-making process, using the plan-do-study-act cycle described by Scherkenbach (1991) as an integrating framework. The planning stage of this cycle consists of four steps: (a) identify current opportunities for improvement, (b) document the present process, (c) create a vision for the improved process, and (d) define the scope of an initial small-scale improvement effort. During the next (doing) stage, a pilot project is carried out in a controlled setting over a reasonably long interval to encounter all relevant obstacles to full-scale implementation. The results are studied carefully during the next stage, leading to an accumulation of knowledge about the unintended as well as the intended consequences of the intervention. Finally, during the action stage, this knowledge is applied to modify the plan by introducing new resources to correct problems identified by the clinical team. The entire cycle is repeated at the next improvement opportunity, resulting in an iterative process that leads to continual improvement if the organizational context includes sufficient resources and a commitment from top-level management to act appropriately on the information uncovered. This data-based approach to decision making is consistent with the goals of applied behavior analysis. Both advocate the development of a plan of action based on a thorough analysis of a problematic situation and the use of corrective feedback provided by the outcomes achieved to make appropriate modifications that are effective in accomplishing the intended objectives.

A troubleshooting guide adapted from Gilbert's (1978) behavior engineering matrix, which was used by Albin (1992) to identify potential obstacles to the implementation of quality improvement efforts in human service organizations, provides a useful framework for considering the types of environmental supports and behavior repertoires that would promote the use of SPC data-analytic and problem-solving tools for clinical decision making in applied settings. In this section, we will consider how some dimensions of Gilbert's behavior engineering matrix may suggest strategies that facilitate the use of control charts in applied settings. *Environmental supports* refer to the characteristics of a work setting that management must create and maintain to accomplish desired outcomes. The three broad categories of environmental supports relate to (a) how information is presented to workers in the form of data, (b) what instruments are available to facilitate the use of control charts, and (c) what motivational conditions have been established to reinforce desired performance. In order to use control charts appropriately, workers must have sufficient data (made available through training manuals and feedback from experienced supervisors about how current performance compares to established exemplars) as well as adequate resources (e.g., materials and equipment) to chart data efficiently. As noted previously, control charts can be done by hand using standardized forms (sample forms are available from SPC Press, Inc., 5908 Toole Drive, Suite C, Knoxville, TN 37919) without a large investment of time and effort. However, choosing the correct control chart for variable or attribute data in a particular application may require the advice of a statistical consultant who is familiar with the assumptions that must be met in order to justify the use of certain mathematical models (e.g., Poisson distributions) to estimate population parameters (measures of central tendency and dispersion) from the statistics computed from the sample data. Professional training is required to set up control charts, but they can usually be maintained by staff members with a high school education. This brings workers into direct contact with clinical data and enables them to share information with other team members about special circumstances that may have been associated with values that are out of statistical control. In this manner, the control chart becomes a living document that embodies the "social memory" of a work group, providing a historical record of the team's decision-making process (see Wheeler, 1986, for an industrial example).

There are three components of a worker's personal repertoire of behavior (knowledge, capacity, and intrinsic motives) that must be present in order for control charts to be used effectively for clinical decision making. The conceptual understanding and technical skills required depend on the complexity of the application and the degree of supervisory support available. Selection of clinically useful target behaviors and measurement strategies requires proficiency in applied behavior analysis (see Johnston & Pennypacker, 1993) as well as experience in SPC. However, after issues about measurement strategies and tactics have been decided, staff members can be given a great deal of responsibility in maintaining control charts for a particular client. This involvement in the decision-making process provides a source of motivation for workers to use control charts that may contribute to pride in workmanship, a principle stressed by Deming (1986, 1993) as important for the successful implementation of SPC in an organization. The use of extrinsic motivation (extra pay, bonuses) to reward workers for using control charts is considered to be counterproductive by Deming because it detracts from this pride of workmanship, although this might be considered to be a questionable assumption by some applied behavior analysts (see Mawhinney, 1992).

The capacity of workers to use control charts in some applied settings can be enhanced by the use of computer programs that handle large databases efficiently. However, some commercially available programs do not compute control limits correctly. Use of computer graphics packages might contribute to a tendency to value the appearance of a control chart over its clinical utility and might alienate workers who do not know how to use these programs. It is important to emphasize that the clinical utility of control charts derives from their ability to help workers distinguish patterns of variation due to special (assignable) causes from variation due to common causes. Having control charts readily available so that staff members can contribute their knowledge of whatever factors may have influenced data points that are out of statistical control by making notations directly on the charts greatly facilitates the ability of the clinical team to respond appropriately. This helps to avoid superstitious decision making, where action is taken on the first data point that appears to be different from others in a time series without first determining if it truly is a signal or merely reflects noise in the system. The orderly accumulation of knowledge that is made possible by use of control charts as part of the plan-do-study-act cycle provides a health care organization with a means for continually improving the quality of services provided to its customers.

In summary, SPC problem-solving strategies that employ cause-and-effect diagrams, scatter plot and histogram analysis, Pareto diagrams, and brainstorming techniques may contribute to the analysis of how various elements in the individual's social ecology contribute to clinical dysfunction in the same manner that these procedures currently facilitate an exploration of the systems variables that contribute to defect production through an analysis of work processes. By systematically including information provided by all staff members who participate in clinical teams, the use of SPC in health care settings may enhance the quality of treatment provided to individuals with challenging behaviors. A similar conclusion was reached by Mawhinney (1987), who examined the application of SPC methodologies in organizational behavior management and noted that "SPC should help us attack more complex performance management systems and reinforcement systems" (p. 4).

REFERENCES

Albin, J. (1992). *Quality improvement in employment and other human services: Managing for quality through change.* Baltimore, MD: Paul Brookes.

Baer, D. M. (1986). In application, frequency is not the only estimate of the probability of behavioral units. In T. Thompson & M.D. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 117–136). Hillsdale, NJ: Erlbaum.

Baer, D. M., & Parsonson, B. S. (1981). Applied changes from steady state: Still a problem in the visual analysis of data. In C. B. Bradshaw, E. Szabadi, & C. B. Lowe (Eds.), *Quantification of steady-state operant behavior* (pp. 273–285). New York: Elsevier.

Bailey, D. B. (1984). Effects of lines of progress and semi-logarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis, 17,* 359–365.

Barrett, B. S., Johnston, J. M., & Pennypacker, H. S. (1986). Behavior: Its units, dimensions, and measurement. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 156–200). New York: Guilford Press.

Birkimer, J. C., & Brown, J. H. (1979). A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis, 12,* 523–533.

Brache, A. P., & Rummler, G. A. (1988). The three levels of quality. *Quality Progress, 21(10),* 46–51.

Cumming, W. W., & Schoenfeld, W. N. (1960). Behavior stability under extended exposure to a time-correlated reinforcement contingency. *Journal of the Experimental Analysis of Behavior, 3,* 71–82.

Deming, W. E. (1986). *Out of the crisis.* Cambridge, MA: MIT Press.

Deming, W. E. (1993). *The new economics for industry, education, and government.* Cambridge, MA: MIT Press.

Fawcett, S. B. (1991). Some values guiding community research and action. *Journal of Applied Behavior Analysis, 24,* 621–636.

Gardner, W. I., & Cole, C. L. (1989). Self-management approaches. In E. Cipani (Ed.), *The treatment of severe behavior disorders* (pp. 19–35). Washington, DC: American Association on Mental Retardation.

Geller, E. S. (1992). Where's the performance in organizational behavior management. *Journal of Applied Behavior Analysis, 25,* 519–523.

Gilbert, T. F. (1978). *Human competence: Engineering worthy performance.* New York: McGraw Hill.

Gitlow, H. S., & Gitlow, S. (1987). *The Deming guide to quality and competitive position.* Engelwood Cliffs, NJ: Prentice-Hall.

Glenn, S. S. (1991). Contingencies and metacontingencies: Relations among behavioral, cultural, and biological evaluation. In P. A. Lamal (Ed.), *Behavioral analysis of societies and cultural practices* (pp. 39–73). Washington, DC: Hemisphere.

Goldiamond, I. (1974). Toward a constructional approach to social problems: Ethical and constitutional issues raised by applied behavior analysis. *Behaviorism, 2,* 1–85.

Groden, G. (1989). A guide for conducting a comprehensive behavioral analysis of a target behavior. *Journal of Behavior Therapy and Experimental Psychiatry, 20,* 163–169.

Haring, T. G., & Kennedy, C. H. (1988). Units of analysis in task-analytic research. *Journal of Applied Behavior Analysis, 21,* 207–215.

Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis, 13,* 543–559.

Hawkins, R. P. (1986). Selection of target behaviors. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 331–383). New York: Guilford Press.

Johnston, J. M., & Pennypacker, H .S. (1993). *Strategies and tactics of human behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kane, V. (1989). *Defect prevention: Use of simple statistical tools.* New York: Marcel Dekker.

Kazdin, A. (1982). *Single case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Killeen, P. R. (1978). Stability criteria. *Journal of the Experimental Analysis of Behavior, 29,* 17–25.

Lindsley, O. R. (1992). Precision teaching: Discoveries and effects. *Journal of Applied Behavior Analysis, 25,* 51–58.

Mahwinney, T. C. (1987). Organizational behavior management and statistical process control: Theory, technology, and research. An introduction. *Journal of Organizational Behavior Management, 9,* 1–4.

Mahwinney, T. C. (1992). Total quality management and organizational behavior management: An integration of continual product improvement. *Journal of Applied Behavior Analysis, 25,* 525–543.

Mainstone, L. E., & Levi, A. S. (1987). Fundamentals of statistical process control. *Journal of Organizational Behavior Management, 9,* 5–21.

Mann, N. R., Charuvastra, V. C., & Murthy, V. K. (1984). A diagnostic tool with important implications for treatment of addiction: Identification of factors underlying relapse and remission time distributions. *Journal of the Addictions, 19,* 25–44.

Montgomery, D. C. (1985). *Statistical quality control.* New York: Wiley.

Moran, J. W., Talbot, R. P., & Benson, R. M. (1990). *A guide to graphical problem-solving processes.* Milwaukee, WI: ASQC Quality Press.

Morris, E. K., & Midgley, B. D. (1990). Some historical and conceptual foundations of ecobehavioral analysis. In S. R. Schroeder (Ed.), *Ecobehavioral analysis and development disabilities: The twenty-first century* (pp. 1–32). New York: Springer-Verlag.

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis* (pp. 157–186). New York: Plenum.

Pfadt, A., Cohen, I. L., Sudhalter, V., Romanczyk, R. G., & Wheeler, D. J. (1992). Applying statistical process control to clinical data: An illustration. *Journal of Applied Behavior Analysis, 25,* 551–560.

Pfadt, A., & Tryon, W. (1983). Issues in the selection and use of mechanical transducers to directly measure motor activity in clinical settings. *Applied Research in Mental Retardation, 4,* 251–270.

Pfadt, A., & Wheeler, D. J. (1993, Summer). Control charts—powerful tools in a clinical setting. *SPC Ink,* 1–5.

Pfadt, A., Wheeler, D. J., Sersen, E., & Moreno, E. M. (1993, May). *Using control charts to evaluate baseline stability.* Paper presented at the 19th annual convention of the Association for Behavior Analysis, Chicago.

Pyzdek, T. (1989). *What every engineer should know about quality control.* New York: Marcel Dekker.

Redmon, W. K. (1992). Opportunities for applied behavior analysis in the total quality movement. *Journal of Applied Behavior Analysis, 25,* 545–550

Rubinstein, S. P. (1984). Quality control requires a social and technical system. *Quality Progress, 17(8),* 22–25.

Saunders, R. R., & Saunders, J. L. (1994). W. Edwards Deming, quality analysis, and total behavior management. *The Behavior Analyst, 17,* 115–125.

Scherkenbach, W. W. (1991). *Deming's road to continual improvement.* Knoxville, TN: SPC Press.

Shewart, W. A. (1986). *Statistical method from the viewpoint of quality control.* New York: Dover Publications. (original work published 1939)

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Skinner, B. F. (1935). The generic nature of the concepts of stimulus and response. *Journal of General Psychology, 12,* 40–65.

Tawney, J. W., & Gast, D. L. (1984). *Single subjects research in special education.* Columbus, OH: Merrill.

Thompson, T., & Zeiler, M. D. (Eds.). (1986.) *Analysis and integration of behavioral units.* Hillsdale, NJ: Erlbaum.

Touchette, P. E., MacDonald, R. F., & Langer, S. N. (1985). A scatter plot for identifying stimulus control of problem behavior. *Journal of Applied Behavior Analysis, 18,* 343–351.

Welch, J. F. (1992). Service quality measurement at

American Express Traveler's Cheque group. *The Quarterly Management Forum, 18(4),* 8–11.
Wheeler, D. J. (1986). *A Japanese control chart.* Knoxville, TN: SPC Press.
Wheeler, D. J. (1990). *Correlated data and control charts.* Knoxville, TN: SPC Press.
Wheeler, D. J., & Chambers, D.S. (1992). *Understanding statistical process control* (2nd ed.). Knoxville, TN: SPC Press.
Wilson, E. B., Jr. (1990). *An introduction to scientific research.* New York: Dover Publications. (original work published 1952)

Received June 16, 1993
Initial editorial decision August 30, 1993
Revisions received January 7, 1994; April 7, 1994; June 16, 1994
Final acceptance June 2, 1995
Action Editor, Richard Winett

APPENDIX

Different formats for recording incident data on the use of protective equipment to control the emotional outbursts of a young woman with severe mental retardation. Duration is the length of time that restraints were used during each shift. Frequency is the number of times that restraints were applied during each shift. Interresponse time is the number of work shifts between incidents.

Date and shift	Duration (in minutes)	Frequency	Inter-response time (IRT)	Instantaneous rate (1/IRT)
April 1 D	85	3	0.33	3.00
			0.33	3.00
			0.33	3.00
E	30	1	1.00	1.00
N	0	0		
	$\Sigma = 115$	$\Sigma = 4$		
April 2 D	0	0		
E	0	0		
N	0	0		
	$\Sigma = 0$	$\Sigma = 0$		
April 3 D	55	1	5.00	0.20
E	0	0		
N	0	0		
	$\Sigma = 55$	$\Sigma = 1$		
April 4 D	30	1	3.0	0.33
E	0	0		
N	0	0		
	$\Sigma = 30$	$\Sigma = 1$		
April 5 D	165	3	2.33	0.43
			0.33	3.00
			0.33	3.00
E	0	0		
N	0	0		
	$\Sigma = 165$	$\Sigma = 3$		
April 6 D	0	0		
E	0	0		
N	0	0		
	$\Sigma = 0$	$\Sigma = 0$		

Date and shift	Duration (in minutes)	Frequency	Inter-response time (IRT)	Instantaneous rate (1/IRT)
April 7 D	170	3	5.33	0.19
			0.33	3.00
			0.33	3.00
E	0	0		
N	0	0		
	$\Sigma = 170$	$\Sigma = 3$		
April 8 D	0	0		
E	0	0		
N	0	0		
	$\Sigma = 0$	$\Sigma = 0$		
April 9 D	250	4	5.25	0.19
			0.25	4.00
			0.25	4.00
			0.25	4.00
E	40	1	1.00	1.00
N	0	0		
	$\Sigma = 290$	$\Sigma = 5$		
April 10 D	65	2	1.50	0.67
			0.50	2.00
E	0	0		
N	0	0		
	$\Sigma = 65$	$\Sigma = 2$		
April 11 D	90	2	2.50	0.40
			0.50	2.00
E	0	0		
N	0	0		
	$\Sigma = 90$	$\Sigma = 2$		
April 12 D	55	1	3.00	0.33
E	0	0		
N	0	0		
	$\Sigma = 55$	$\Sigma = 1$		

APPENDIX
(*Continued*)

Date and shift	Duration (in minutes)	Frequency	Inter-response time (IRT)	Instantaneous rate (1/IRT)	Date and shift	Duration (in minutes)	Frequency	Inter-response time (IRT)	Instantaneous rate (1/IRT)
April 13 D	0	0			April 22 D	90	2	4.50	0.22
E	0	0						0.50	2.00
N	0	0			E	0	0		
	$\Sigma = 0$	$\Sigma = 0$			N	0	0		
						$\Sigma = 90$	$\Sigma = 2$		
April 14 D	100	2	5.00	0.18	April 23 D	40	1	3.00	0.33
			0.50	2.00	E	0	0		
E	0	0			N	0	0		
N	0	0				$\Sigma = 40$	$\Sigma = 1$		
	$\Sigma = 100$	$\Sigma = 2$							
April 15 D	0	0			April 24 D	155	3	2.33	0.33
E	30	1	4.00	0.25				0.33	3.00
N	0	0						0.33	3.00
	$\Sigma = 30$	$\Sigma = 1$			E	0	0		
					N	0	0		
						$\Sigma = 155$	$\Sigma = 3$		
April 16 D	0	0			April 25 D	140	3	2.33	0.43
E	0	0						0.33	3.00
N	0	0						0.33	3.00
	$\Sigma = 0$	$\Sigma = 0$			E	0	0		
					N	0	0		
						$\Sigma = 140$	$\Sigma = 3$		
April 17 D	0	0			April 26 D	220	4	2.25	0.44
E	0	0						0.25	4.00
N	0	0						0.25	4.00
	$\Sigma = 0$	$\Sigma = 0$						0.25	4.00
					E	45	2	0.50	2.00
								0.50	2.00
					N	0	0		
						$\Sigma = 265$	$\Sigma = 6$		
April 18 D	0	0			April 27 D	0	0		
E	0	0			E	0	0		
N	0	0			N	0	0		
	$\Sigma = 0$	$\Sigma = 0$				$\Sigma = 0$	$\Sigma = 0$		
April 19 D	0	0			April 28 D	0	0		
E	0	0			E	30	1		
N	0	0			N	0	0		
	$\Sigma = 0$	$\Sigma = 0$				$\Sigma = 30$	$\Sigma = 1$		
April 20 D	85	2	13.50	0.07	April 29 D	78	2	1.50	0.67
			0.50	2.00				0.50	2.00
E	70	2	0.50	2.00	E	90	2	0.50	2.00
			0.50	2.00				0.50	2.00
N	0	0			N	0	0	0.50	
	$\Sigma = 155$	$\Sigma = 4$				$\Sigma = 168$	$\Sigma = 4$		
April 21 D	0	0			April 30 D	30	1	2.00	0.50
E	0	0			E	0	0		
N	0	0			N	0	0		
	$\Sigma = 0$	$\Sigma = 0$				$\Sigma = 30$	$\Sigma = 1$		