Decision making with computer graphics in an inventory control environment

by J. S. PROKOP
Office of the Secretary of Defense
Washington, D.C.
and
F. P. BROOKS, JR.
University of North Carolina
Chapel Hill, North Carolina

INTRODUCTION

Computer-driven displays have long been thought to help decision making. But the justification for using these devices in decision-making has been long on intuition and short on quantitative analysis. To see if this intuition was right, we conducted an experiment.

Eighteen interested and experienced decision-makers in the inventory control field met for twenty class hours of instruction in advanced inventory control techniques. Near the conclusion of the short course, we measured the participants' decision-making ability while using a computer-driven display device. This measurement was compared to their decision-making ability while using information presented to them on paper.

To provide a vehicle for the investigation, a simulator was written to apply certain inventory control policies to a hypothetical inventory system handling 34 items. This inventory system faces a randomly derived set of orders, price changes, replenishment of stock and other transactions. The simulator has two sets of input parameters: one governs the distributions of transactions; the other establishes the management policy for inventory control. The simulator investigates twelve distinct policies at one time over a simulation cycle of 24 months.

Each participant was given, as examination problems, the statistics resulting from two different simulation runs. Each decision-maker, or decider, made a series of decisions; at the end of each simulation month, he ranked the twelve policies in order of desirability. Each decider used printer output on one problem and graphic display presentation on the other.

The statistics from the simulation program were displayed in a variety of formats on standard computer printer paper. The statistics included:

1. percent availability of stock,
2. number of purchase orders generated,
3. lost sales,
4. total dollar investment in inventory.

The output was in the form of tabular listings and bar graphs. Also, for each problem, an IBM 2250 cathode ray tube display unit was programmed to show the same data as appear on the printer output, in fundamentally the same listing and graphical formats. Thus the experiment did not attempt to establish the relative comprehensibility of listed versus plotted data. Instead it investigated the relative comprehensibility of printed versus display presentation of such data, both printed and plotted. The display unit was under the control of the decider, who was able to specify which of the twenty-four months and the kind of displayed data which he wanted to see. In both display and printer output cases, he was expected to start at the first month and proceed through the months in order, with freedom to review.

The course participants were divided into two test groups, each group consisting of nine individuals chosen at random. Tests were made to answer:

1. Are decisions made earlier? The simulated calendar month (out of the twenty-four months of simulation) in which the decider feels he has enough data to commit himself to a ranking for future action was tested.
2. Are decisions made faster? The elapsed clock time to decide on a ranking of policies which the decider would be willing to commit himself to for
future action was tested. The elapsed clock time to complete the remainder of the problem was also tested. After committing himself to a ranking, the participant continued the problem through the twenty-fourth month.

3. Are decisions made more consistently? That is, does the decision made at the month of commitment agree better with that made at the twenty-fourth month, after all data are known?

4. To what degree do the members of each group agree among themselves in regard to the rankings?

In summary, the decider made two basic kinds of decisions. He decided on a ranking at each month in turn, and he decided whether or not to indicate at this month that he felt confident enough of his ranking to recommend that it be used for governing action.

The resulting ordered lists of policy sets and the preparation times were analyzed in two basic ways. First, a 2 x 2 Latin square design with repeated measures was used in an analysis of variance. Rank order statistics were then used to test the consistency of each individual decision-maker and the concordance of a group of decision-makers.

The statistical analysis of variance allocated the observed differences into (1) differences between graphic and printed data presentation (treatment), (2) differences between the first and second problems done by each participant (order), and (3) differences between the performance of the two groups the class was divided into.

The ranking of the policies at the decision month and the ranking at the end of the entire twenty-four month simulation were compared by means of the Spearman rank correlation coefficient. A Kendall’s coefficient of concordance was computed to measure how well the rankings produced by each test group on each problem agreed among themselves.

The results may be summarized as follows:

1. The time to make a decision was decidedly reduced in favor of the display presentation.
2. A decision could be made on the basis of less information by using the display device; that is, action decisions were made earlier in the simulated two years.
3. More consistent decisions were arrived at by using this display device. Even though action decisions were made earlier, with less information, using the display device, the decisions made matched better those made with all data available.

THE PRESENTATION TO THE DECISION-MAKER

The simulation statistics for one month are shown in Figures 1, 2, and 3. The total investment figure is scaled by 100,000. A total investment of $423,000 is, therefore, represented as $4.23.

Figure 2 is a table which is ordered by several of the statistics of interest. Figure 3 shows one statistic across all policies as a bar graph. The printed bar graphs are produced monthly for availability, value of lost sales, purchase order activity, and total investment.
Once the printer output was specified, these presentations were mimicked on the IBM 2250 cathode ray tube. Some exploration of the power and utility of such a display device was made while maintaining similar presentation formats for sake of the experimental design. Figure 4 is the display device representation of Figure 2, for example, and shows some rearrangement of material due to display character and line number limitations. Figure 5 shows the first marked departure from the printer output of Figure 3. The layout is essentially the same, but the presentation is dynamic. A programmed timer advances all lines of the bar graph simultaneously at one second intervals. If the user specifies a month at which he wishes to examine the data, and then specifies the bar graph he wishes to see, the graph begins at month 1 and then advances at one-second intervals to the month requested. The user may alternatively specify half-second intervals. The movement of the bar graph gives a feeling for the history and current derivative of the statistic under scrutiny. In addition to the movement of the bar graph, the symbols <, > and = are used on the display unit to indicate whether the value of a particular bar has increased, decreased or remained equal in comparison with the previous month.

Figure 6 has no printed correspondent in this experiment. While the data for cost of lost sales, availability, total dollar investment and purchase action activity are available as individual bar graphs as represented by Figure 5, all four of these are available.
at the same time in the Quadrant Graph. All bars move under the same program timer control as the individual graphs.

The user indicates by a Programmed Function Keyboard (PFK) to the computer what he wishes to see on the display device. The user is allowed to refer to any past month's data. He may not, however, move ahead in simulated time by more than one month at a time.

THE SUBJECTS OF THE EXPERIMENT

The experiment was designed to examine the decision-making processes of experienced, practical and interested professional administrators of inventory control. Hence the subjects were solicited carefully. A short course in advanced inventory control techniques was designed as a graduate-level course which would attract only those who were interested in the subject matter and who were prepared to understand the material. The simulation model was the experimental vehicle and was used in examples and problems which were integrated into the course. The participants came to the course in statistical inventory control for their own professional advancement, under the auspices of the firms for which they worked. The participants were not informed of the experiment which was being conducted, and they received the full measure of instruction for which they came.

We solicited participants from nearby manufacturing companies specifically for this short course. Candidates for the course had to have experience in inventory control and had to be in a decision-making position in the company. No company was allowed to sponsor more than two candidates.

The participants attended the course for two hours every weekday morning for two weeks. The mornings were chosen specifically to ensure that the attendees would have the active support of their sponsoring firms.

The subjects were divided randomly into two groups of nine people each. After a substantial part of the course material had been given, the participants were given either a problem book with twenty-four months of simulation output (Figures 1, 2, and 3) or assigned to work a problem on the display unit at individual laboratory sessions. The second problem was assigned later in a similar manner except the groups now used the presentation method that they had not used on the first problem. It was explained that congestion on the display unit precluded everyone working both problems in this way. There was no attempt made to identify individuals by group membership or emphasize this distinction.

The participants were not graded on these problems and were encouraged to work them as part of the education process of the course to investigate some of the fundamental properties of the inventory control policies. Although all of the inventory control policies simulated for the two problems were discussed in class, the specific problem policies were not identified, in order to avoid obvious bias. The participants, then, were asked to make their judgment based solely on the simulation results. The two problems were of equivalent difficulty.

An operational setting was postulated for the problem. Higher-level management had presented the participant with the output data for the set of twelve inventory control policies, and had requested a recommended ranking so that an implementation of policies could be decided upon. The participant was encouraged by management to present his recommended ranking as soon as possible, but cautioned that implementation of the recommended policies could have serious repercussions in the firm, so his best professional judgment was required. Participants were reminded to keep accurate figures on elapsed time spent studying each month's data, and were also reminded to mark the decision month at which they would have presented that ranking as an action recommendation to higher level management.

In such an experiment there are two possible approaches to the problem of evaluating the quality of the decisions made. The most commonly used approach is to furnish the subjects with the criteria which are to be used in evaluating the decisions. For example, in many cases the subjects are first given instruction, then tested to ascertain how well they perform as measured against the pre-established criteria.

This method of measurement may be dependable when the material is objective and the criteria are easily established. In inventory control (and many other problems) the weightings to be used in reaching a decision are highly individual. For example, one firm may emphasize high availability, whereas another may give heavy weight to low investment. When the experimental subjects are experienced in the field and have developed their own criteria for decisions, the method of measuring against instructor-set criteria is unsuitable. One cannot know the extent to which the subjects followed the instructor's criteria and the extent to which they followed their own.

In this experiment we followed an alternative approach to the problem of evaluating the quality of decisions. The subjects were experienced in decision-making and were familiar with and interested in the substance of the course. Therefore, they were given no weightings, no criteria of excellence by the instructor. Instead they were explicitly told to apply their own several diverse sets of criteria. The quality of their
decisions was then measured by the consistency of each subject's own ranking at the month of commitment with the later ranking when full information was available.

Although this second approach almost eliminates the meaningfulness of comparisons which show how well the subjects agree with each other, such measures were taken as a matter of interest. The use of self-consistency rather than artificial criteria does, however, improve the credibility of the other statistics. It also substantially improves the generalizability of the results, for it shows the effect on decision-makers when deciding by their own standards, rather than by those dictated by a simplified theoretical model. In substance, one attempts thereby to examine the decision process as it works under operational, rather than classroom, constraints.

THE DESIGN OF THE EXPERIMENT

Two main analyses were performed on the data: an analysis of variance of the Latin square design, and a computation of rank order statistics.

In the Latin square design of this experiment, it should be noted that instead of experimenting on different subjects in each cell of the square, the same experiment group was involved in both cells of a row. In effect, each group acted as its own control group. Experiments in which the same subjects are used under all treatments require q observations on each subject, and are called experiments with repeated measures. In this experiment q = 2 since we are dealing with two treatments and each subject is observed twice.

The 2 x 2 Latin square with repeated measures of Table I was used to analyze the data for:

1. Simulated calendar time to reach a decision (decision month),
2. Actual elapsed time to complete the rankings through the decision month,
3. Actual elapsed time to complete the problem after the decision month,
4. Correlation coefficients for the consistency of the committed decision with the final ranking.

<table>
<thead>
<tr>
<th>Group I</th>
<th>Order 1</th>
<th>Order 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(9 Individuals)</td>
<td>Display</td>
<td>Printer</td>
</tr>
<tr>
<td>Group II</td>
<td>(9 Individuals)</td>
<td>Printer</td>
</tr>
</tbody>
</table>

Table 1---Latin Square Design

in order to produce an analysis of variance of these statistics. When we discuss the effects of treatment in the following pages, we will be discussing whether the display device or the printer demonstrated more influence in our measurements. The effects of order will refer to the influence of the participants’ doing the problems in the order given to them.

The particular design was chosen in part to isolate that variation due to natural differences between the experimental groups. The separation of the source of variation due to group differences allows a better measure of that variation due to the form of data presentation, which is the statistic of real interest.

Since the values of the decision month are positive integers, they were, as is standard, transformed by taking the square root, to make the variances more homogeneous and at the same time to normalize the distribution. It is this square root value which is used in the analysis of variance computations, and reported in the results.

It is of interest to compare the rank order of the inventory control policies at the decision month and the rank order at the twenty-fourth month for each individual. This comparison gives a measure of each individual’s consistency between the ranking decision made at the decision month and that ranking decision made when full information was available. These two rankings for each participant are compared by means of a Spearman rank correlation coefficient (rs) test. Once computed, these values for rs are treated as data for analysis in the Latin square design previously discussed. Since the rs values are known to have a skewed distribution, a standard transformation was made on each of the rs values in order to place the data on a symmetric scale, so as to normalize the distribution.

In order to find the extent to which the members of each group ranked the policies the same way under the same conditions, Kendall’s W coefficient of concordance was computed for each cell of the Latin square. From this, an average rs value of Spearman’s coefficient was computed for each of the four cells. This gave a measure of the homogeneity of decisions under the conditions of each cell. These values of average rs indicate a degree of concordance, or how well the group members agreed among themselves as to the rankings. In the interpretation of the result of this computation it should be noted that this experimental group was not attempting to apply a common criterion of excellence in making their decisions. Individual best judgment and experience in the decision process guided the problem solution. The course which the subjects attended neither taught nor encouraged uniformity in decision-making or in performance goals.

The difference between the two values of rs just
discussed is that the single value of average $r$, derived by way of Kendall's $W$ statistic is a measure of agreement among all nine participants within a cell; whereas the individual values of $r$, derived from Spearman's test are a measure of each person's consistency.

RESULTS

The data analyzed and the detailed results of the analyses are available from the authors. This section of the paper will highlight some of the results.

The usual method in statistical hypotheses testing involves setting the significance level of the test in advance of obtaining the data. The convention used in the analysis of variance deviates somewhat from this formality. The value of $F$ (the significance level of the variance under investigation) is reported to exceed a specified percentile by a comparison with tables of critical values. This allows each reader to establish his own significance level and to judge the results thereby. We will consider a conclusion to be more surely established if the probability of its truth is higher. This methodology does not allow a measure of the power of the test; however, the procedure is valid for estimating the probabilities of the observation in relationship to the assumed sampling distribution. In this investigation we would have been encouraged to find significance at the .05 level, and the significance levels actually attained are noted in this discussion.

The decision month

In this analysis of variance, we were principally interested in investigating the effect of the treatment on the subjects. Our first question was: How early in simulated time can a decision be made?

We found that there was less than one percent probability that the observed differences between treatments were due to chance for this question. We can say with a high degree of assurance that the mean decision month arrived at by using the display unit was indeed less than that using the printer output. From another viewpoint, the participants on the average needed to look at less data with the display unit to make a committing decision than they did with the printer output. No effect could be attributed to order.

The time to make a decision

The elapsed clock time for a participant to commit himself to a ranking was analyzed next. The mean time to decision using the display unit was, with great certainty (> .999), less than the mean time to decision using the printer output. This indicates that the amount of time spent in making a decision was significantly reduced in this experiment by using the display device, confirming the intuitive belief that a person can assimilate a large quantity of data and correlate these data by using display techniques, as opposed to printer output. The convenience of having virtually instantaneous recall of data displays by using the Programmed Function Keyboard is certainly a consideration in the interpretation of the results. Pushing buttons is just inherently faster than paging through a book of data, however well arranged and indexed the book may be. However much or little this consideration affected the results, they show that data can be correlated faster and retained better from a properly programmed display unit.

Unsolicited comment from individual participants supported this conjecture without exception. We observed that the dynamic graphs gave the participant a much better intuitive feel for the situation, and that he was more likely to retain this impression and not have to refer to past data repeatedly. A major help seems to have been the program control which always started the dynamic graphs at month 1 and brought them up to the current month in increments of one month per second. This forced a continual review of the history and derivative of the measure under consideration and undoubtedly reinforced past impressions. It was seldom that a participant asked that the graph be stopped at a month prior to his current month so that he could review the static situation as of that past instant. We noted when the experiment was well under way that the participants with more experience as inventory managers used the dynamic graphs extensively, where the less experienced participants relied on the tabular listings presented on the display unit.

The time to finish the problem

The analysis of variance of the total elapsed time to complete the problem after the decision month was examined. One would expect some speed-up by participants after they had made their commitment decision, simply to get to the end to see how well they did. There is some evidence of this speed-up in the time data. How much of this is due to increased familiarity with the problem at hand and how much is due to impatience to get to the final result is difficult to say. The elapsed time after the decision month was tempered by the requirement that the data be ranked at each month. From personal observation, the participants appeared to be conscientious about following the spirit and the letter
of the instructions, but relieved that the big decision had been reached, and were in a hurry to finish the twenty-four months to check their final ranking against their decision month ranking.

The order factor was significant at the .01 level, which is reasonable and consistent with our previous comments on the effects of order. In this test the treatment factor was completely without significance which is also a reasonable result in view of the observation concerning the impatience to finish the problem.

**Individual consistency in decisions**

The transformed Spearman rank correlation coefficients, $r_s$, were then subjected to an analysis of variance. We found a significance level of .05 for the treatment, and we obviously do not have the clear mandate that our other treatment factor statistics have given, but the evidence is that the mean correlation coefficient is higher using the display unit than using the printer output. The values of $r_s$ give a correspondence between the participant's ranking of the policies at the decision month and his ranking at the last month of the simulation data, month 24. This, then, is a measure of the consistency between these two rankings. It is also a measure of the participant's discrimination ability—that is his ability to decide whether he has enough information to commit himself to a ranking or not. A decision to commit too soon in relation to each individual's ability and ranking criteria would, in most cases, result in a poor correlation coefficient, whereas being overly cautious and waiting beyond the point where he had sufficient information could not be expected to materially improve the correlation coefficient. Thus, we might say that one interpretation of a low $r_s$ would be that the participant committed himself too early. Other interpretations are, of course, that he simply used poor judgment in his ranking, or that he materially changed his ranking criteria after the decision month. Participants were cautioned to use a consistent ranking schema throughout. As an extreme example, we pointed out to the class that to rank the policies based only on lost sales data through the decision month, then to abandon that schema and to rank the policies only on number of purchase orders generated would not be showing responsible judgment. On the basis of these comments, we should be able to narrow our consideration of a principal cause of low $r_s$ to either too early commitment or poor ranking judgment at some point. Both of these essentially are measures of decision-making ability and we can accept either one or both as reasonable interpretations of a low correlation coefficient. Observe that the greater consistency observed for the display results is in spite of the earlier month of commitment. This earlier commitment would be expected to make consistency worse.

For the transformed Spearman correlation coefficient, $r_s$, the order factor was highly significant; the significance was at the .001 level. Thus we accept the hypothesis that maturation would appear to play a large role in the decision-making consistency that is being measured. That is, the ability to meaningfully rank a set of policies grows with practice.

**Group consistency in decisions**

The results of the decision process at the end of the twenty-fourth month, when the decider had full information available to him, were next examined for average $r_s$. The resulting average $r_s$ is represented in our usual Latin square format in Table II. Here the effect of Group II going from printer output at Order 1 to display unit output at Order 2 is pronounced. There was a moderate increase in the average $r_s$ for Group I going from display to printer output, which may be ascribed in part to maturation. However, the average $r_s$ almost doubling in Group II when going from printer output to display output may be more than can plausibly be ascribed to maturation alone. In the case of these rankings at the twenty-fourth month, the values of average $r_s$ are a valid measure of concordance. Remember that in the rankings at the twenty-fourth month, the deciders all had the same amount of information available to them, which would not have been the case for a decision month ranking. Additionally, the participants at this point were concerned only with the ranking process, and not the additional question of whether or not to indicate a decision month.

**CONCLUSIONS AND REMARKS**

The strongest result statistically was that actual time to make a decision was reduced in favor of the display presentation.
A second statistically significant result was that a decision could be made earlier, or on less complete data, by using the display unit. This is the most significant result economically. And a one percent result is very strong, statistically.

The mean time over both problems to reach the decision month using the display device was 52 minutes. The mean time using printer output was 82 minutes. This result points to the use of a display presentation when economy of time or simply volume of printed output is a serious constraint on the system or the decision-maker. The mean decision month was 9.2 while using the display device and 11.3 for the printer output.

While the two results reported above have the respectability of high statistical significance, the next results to be discussed are at least as important in the evaluation of the experiment. These are the results which answer the question of whether a better decision can be made with a display device. We will claim that a decision at the month of commitment that is more consistent with the final decision is a better decision.

The results from Kendall's W statistic and from Spearman's rank correlation statistic show that the subjects when using the display tended to make decisions more consistent with themselves, and even with their group.

The economics of a system of display devices for decision-making will not be explored here, however, it is evident that the very specialized research equipment used for this experiment is both expensive and unnecessarily elaborate for an operational system.

The minimum display unit for implementation of an information system of this general type should have alphanumeric display capabilities and a programmed function keyboard, or equivalent means of easy display selection. The size of the display face is crucial to the extent that it must be able to contain enough material to be of interest and still allow character size and spacing to enhance readability. For instance, the IBM 2250 used in this experiment has a display face twelve inches square with a maximum capability of fifty-two lines of seventy-four characters each. The information in the displays for this investigation is rather densely packed and is digestable only by someone sitting in the console chair immediately in front of the display face. A smaller display face would mean that displays would have to be segmented; the same information in smaller characters on a smaller display face would be the wrong compromise. With segmented displays, more programmed function keys would be needed and the problem of how to ask for a particular display becomes more complicated for the user. It is unfortunate that the great majority of available alphanumeric display units have small display faces—eight inches by six inches appears to be a popular size. Other features of the IBM 2250, such as an alphanumeric typewriter keyboard, line-drawing capability, and a light pen, are unnecessary for this application.

In addition to the display device proper, this experiment used other system facilities. The display unit had a self-contained buffer of 8,192 characters. Of this, a maximum of 2,000 characters of buffer storage were used at any one time. The display program in the main storage of the IBM 360 Model 40 used approximately 13,000 characters for program and 21,000 characters for tables. An additional 46,000 characters of disk storage were used for table overlays.

The display system evidently achieved the objective of presenting a complex situation, which involved many inter-relationships, in a manner such that the key concepts and fundamental correlations were clearly understandable. The display system appeared to facilitate interpretation and extrapolation of the relevant data. The reduction of reaction time of top-level decision-makers in this environment is both an interesting result and an important objective of any executive display system.

One point of great interest for further work would be the exploration of the differential cost or savings of decisions using display units and printer output. This is a rather difficult area to define, since dollar values and weightings must be assigned not only to the reduction in inventory valuation and the cost of lost sales, but also to the generated purchase actions, availability of material, timeliness of decision, system cost and other factors.

Statistics on the frequency of use of the various displays should be collected, both automatically and by experimenter observation. The correlation of the frequency of use by display type with the individual's consistency of decision would be most important for the design of extensions of this system.

Whatever the extension of the experiment, there should be the capability for the decider to request a hard copy of any display he wishes. If line drawing capability is used, this, of course, implies the availability of the equivalent of a line plotter for hard copy output. This requirement is more operational than experimental. We have no doubt as to the utility of such a feature for the decider in an operational environment, and if a display unit has a line drawing capability, it should be used with this requirement in mind.

SELECTED BIBLIOGRAPHY

*Computer characteristics quarterly (second quarter)*
Keydata and Adams Associates Inc Watertown Massachusetts 1968
R G BROWN
Statistical forecasting for inventory control
D S BURDICK  T H NAYLOR
Design of computer simulation experiments for industrial systems
Communications of the ACM 9 May 1966 pp 329-339
W G COCHRAN  G M COX
Experimental designs
John Wiley & Sons, Inc New York 1962
W L HAYS
Statistics for psychologists
Holt, Rinehart and Winston New York 1963
IBM system/360 operating system graphic programming services
for IBM 2250 display unit
International Business Machines Corporation Form C27-6999-4
December 1967
IBM system/360 component description, IBM 2250 display unit
model 1
Form A27-2701-1 January 27 1967
H R LUXENBERG  R L KUEHN eds
Display systems engineering
T H NAYLOR  J L BALINTFY  D S BURDICK  C KONG
Computer simulation techniques
John Wiley & Sons New York 1966
K WERTH  T H WONNACOTT
Methods for analyzing data from computer simulation experiments
Communications of the ACM 10 pp 703-710 November 1967
G W PLOSSL  O W WIGHT
Production and inventory control
Prentice-Hall Inc Englewood Cliffs NJ 1967
H SACKMAN  W J ERIKSON  E E GRANT
Exploratory experimental studies comparing online and offline
programming performance
Communications of the ACM 11 pp 3-11 January 1968
R G D STEEL  J H TORRIE
Principles and procedures of statistics
B J WINER
Statistical principles in experimental design