M. N. Crook*
D. S. Kellogg

Experimental Study of Human Factors for a Handwritten Numeral Reader

This Letter reports a preliminary study of the effects of various human factors on the man-machine interface for an experimental handwritten numeral reader.[1] Previous papers[2,3] have reported results of similar studies for a machine which read numbers written under a two-dot restraint. The objectives of these earlier experiments were to determine the feasibility of acceptable performance and to evaluate motivational factors. The numeral reader used in the present experiment utilizes a contour analysis technology that allows the numerals to be written with considerably more freedom in registration, size and shape. The scope of the investigation we are reporting includes training methods, physical conditions, mental stresses, effects of incentives, variations in writing equipment and population factors.

The experiment was conducted, for the most part, at the Institute for Psychological Research at Tufts University. Subjects were assembled, classroom style, in groups of approximately ten, and were given a variety of instructional material and writing assignments for about one hour. Machine results were usually available for feedback to the subjects. Sources included the student body at the University, the student body from local high schools, some secretarial personnel, sales personnel from a large metropolitan department store, and clerical personnel from a statistical analysis group. The subjects were usually paid for their time. When incentives were being explored, subjects were also given an opportunity to earn bonuses for good performance.

The subjects were asked to write, under controlled conditions, sets of numerals, which were then read by the experimental machine. Certain numerals were not machine readable (rejects), and others were incorrectly read (substitutions or errors). For purposes of evaluation, the reject rate for each subject was taken as the quantitative measure of human performance. Error rates were too low to be useful for this purpose.

For the particular recognition logic used during the experiment, it was important that the subjects write numerals with consideration for shape factors, such as

1) Gaps (as in top of **5**) were not permitted.
2) Bays (as in **2, 3, 5**) were required to be open.
3) Loops (**6, 8, 9, 0**) were required to be well rounded and enclosed.
4) Lines could not cross over materially where they closed a figure (at top of **8**'s and **0**'s).
5) Fancy strokes and extra-long tails were to be avoided.
6) Numerals were to be well proportioned, with proper balance between upper and lower portions.

Within the above limits, there was considerable latitude for shape variation.[4]

For each writing condition to be analyzed (with minor exceptions), the subjects were asked to write 100 numerals, 25 on each of four IBM® cards that were designed specifically for the handwritten samples. The sequence of events that included instructing a group of subjects, having the subjects write 100 numerals on four cards, and reading these cards in the experimental machine was known as a *cycle*. A sequence of cycles, during which the same subjects were being tested without recess, was known as a *period*. In most cases, subjects were asked to return for a second period. Two periods would usually consist of eight to ten cycles.

The statistical and sales personnel were tested off campus for three cycles, with a delay between the second and third cycles.

In the usual routine, subjects copied numerals from lists, working at their own pace. In some periods, a light mental work load was imposed by requiring the subjects to add successive rows of the numerals when copying them onto the cards. The effect of a time limit was tested with some groups by giving the subjects an ample supply of cards and asking them to write as many as they could in either five or six minutes. There were 223 subjects; of these, 120 were college students, 52 high school students, 29 sales clerks, and 22 statistical clerks and miscellaneous. The subjects served in 21 experimental groups. Of these, 15 groups served for two experimental periods.

Raw writing data, without instructions about writing for the machine, were obtained from 181 of the subjects.

Some information has been obtained on about 30

*Tufts University.

variables in the experiments. In some cases, statistical evaluation was possible. Since it was impractical to apply rigorous controls in every experiment, some evaluations were based on the experimenter's judgments and impressions. Performance was evaluated primarily in terms of quality rather than speed.

Experimental results

The reject rates for the individual variables are given below for comparative purposes only; they do not necessarily represent mean performance levels, after appropriate training, for larger samples of the population. Only those variables which are significant and/or are of special interest have been included. Except where otherwise indicated, data are obtained from the student populations only. All statistical evaluations of mean differences were made by means of *t*-tests.

- *Training*

The effects of training (after training methods had been refined) are shown in Fig. 1. Cycle 1 results represent raw scores, before training. The reject rate falls sharply after the first training period, and continues to fall with additional practice. The curve for substitution rate has an almost identical shape, the substitution ordinate scale being 1/10th the reject ordinate scale.

If subjects are ranked on the basis of percentage of substitutions, they produce the type of results shown in Fig. 2. As shown in the Figure, 15% of the subjects are responsible for approximately half of the substitution errors.

- *Work Pressure*

Arithmetic. Results obtained with characters written when the subjects were doing simple arithmetic were compared with results obtained with characters written without arithmetic. Four comparisons were made using characters written by the same group of subjects. In each case slightly higher rejection rates occurred on the arithmetic cycles, but in only one of these comparisons was the difference statistically significant (at the 0.05 level). When the four cases were averaged, the mean rejection rate for the non-arithmetic cycles was 0.9% and for the arithmetic cycles, 1.4%.

Time Limit. Results for timed and untimed cycles were compared using data from the same subjects. The mean untimed rejection rate was 1.0% and the mean timed rejection rate, 2.0%. This difference was not statistically significant.

Figure 1 **Composite learning curve for 72 experimental subjects. "Cycle" is defined in the text.**

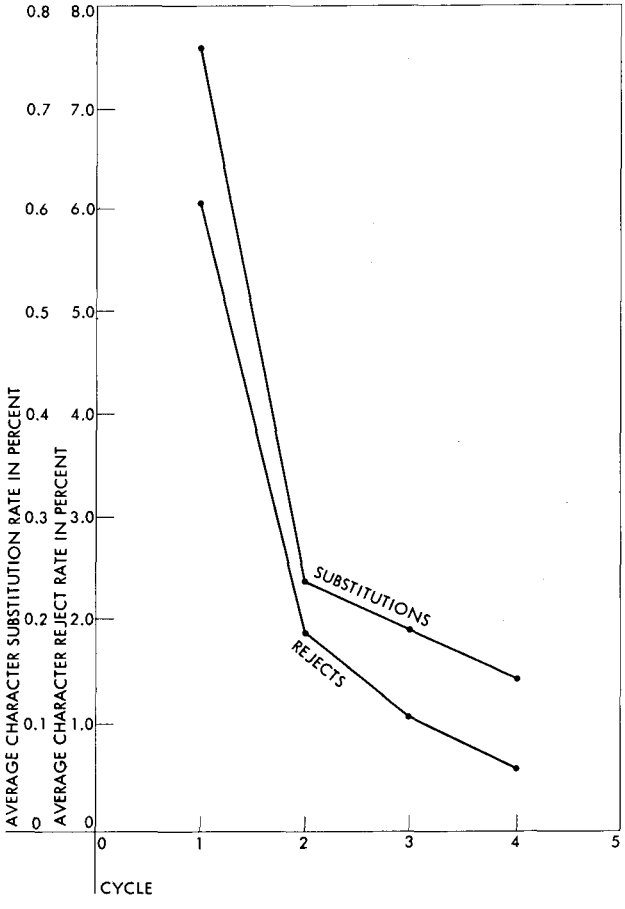
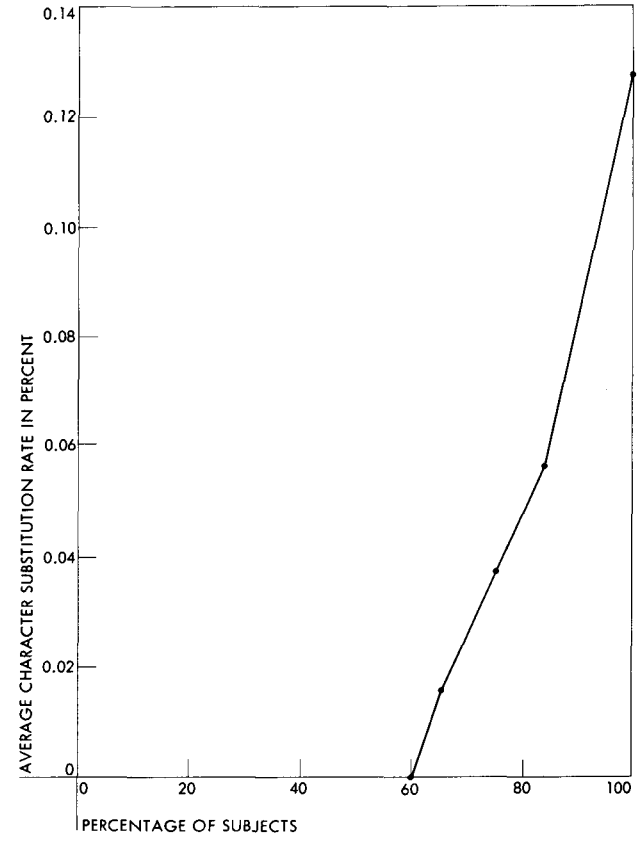


Figure 2 **Relation of the error rate to the percentage of trained subjects.**



- *Incentives*

The incentives were judged to be effective, although the interaction with other conditions prevented a rigorous statistical evaluation. They appeared to bring down the final level of scores, to minimize boredom, to reduce erratic or careless performance, and to facilitate learning in the early stages.

- *Population factors*

For the population analysis, two types of rejection scores were used: 1) Percent rejection in the first cycle, based on first cycles which produced a reasonable approximation to raw handwriting performance; and 2) percent rejection in one or more late cycles, uncontaminated by special experimental conditions.

Educational level

The mean rejection rate for college subjects in Cycle 1 was 10.1% with a PE (probable error) of 0.51 (N = 102). The corresponding figure for high school subjects was 8.4% with a PE of 0.89 (N = 32). The difference is not significant. In late cycles, the mean rejection rate for college subjects was 1.59% with a PE of 0.13 (N = 118), and for high school subjects, 1.04% with a PE of 0.17 (N = 45). This difference again is not significant.

Clerical personnel

The performance level of the sample of statistical clerks was very similar to that of the high school and college subjects in the latter part of the program when incentives and modified training procedures were in effect. Mean third-cycle score for the ten clerks was 1.2%; the corresponding score for high school and college subjects combined was 1.4%. The difference is not statistically significant.

Sales clerks

The performance level of a group of 27 department store sales clerks was poorer than that of the high school and college subjects. The mean rejection score on the third cycle was 2.6%; this can be compared with 1.4% for the subjects cited above. The difference is statistically significant at the 0.05 level. This comparison, combined with the experimenter's impression that the subjects had more difficulty understanding the instructions which had been evolved for the student subjects, suggests very strongly that the sales clerks represent a different population, with respect to the function being studied, from that represented by the college and high school subjects.

The sales clerks were then divided into a group of ten who had been in the job less than 11 years and a group of 16 who had been in the job 11 years or more. (One subject was omitted because of unclear record.) The mean third-cycle score for the relatively short-term group was 1.3% and for the relatively long-term group 3.5%. This difference is significant at the 0.01 level, and suggests a real difference in the population. These results imply that age per se is not, however, the critical factor.

Sex

In Cycle 1, the mean rejection rate for the females was 9.6% with a PE of 0.65 (N = 62); the corresponding score for males was 9.8% with a PE of 0.61 (N = 72). In late cycles, mean score for females was 1.2% with a PE of 0.12 (N = 76), and for males, 1.6% with a PE of 0.16 (N = 87). These differences are not significant. A further breakdown of the data by both school and sex categories disclosed one comparison of interest: the difference between male and female high school subjects in late cycles. The mean for females was 0.4% with a PE of 0.07 (N = 25), and for males, 1.8% with a PE of 0.34 (N = 20). This difference is significant at the 0.02 level.

Although a similar difference was not apparent in the full sample comparisons for the college subjects, examination of subgroups within the same experimental periods yielded data which, though numerically of little weight, support the sex differences found in the total high school sample.

The data, though not statistically conclusive, make it unwise to dismiss the possibility of sex differences of moderate magnitude in some population groups.

Conclusions

On the basis of the testing and analysis conducted thus far in the study, the following conclusions are drawn:

1) People can modify their writing habits to a significant extent with the expenditure of approximately 30 minutes of training effort.
2) The effects of imposed physical conditions and work pressures, simulating some typical work situations, have not been shown to produce significant deterioration of performance where proper motivation exists.
3) Population factors may be significant in some areas of application, making specialized training methods necessary. In addition, some personnel selection may be required, since approximately 15% of the personnel account for about half of the substitutions errors.

References

1. E. C. Greanias, et al., "The Recognition of Handwritten Numerals by Contour Analysis," *IBM Journal*, this issue, p. 14.
2. R. S. Hirsch, et al., "Experimental Evaluation and Statistical Analysis of Human Performance in a Constrained Handwriting Method of Data Recording," IBM ASDD Technical Report 16.10.070.019, June 1, 1960.
3. J. L. Masterson and R. S. Hirsch, "Machine Recognition of Constrained Handwritten Arabic Numbers," *IRE Transactions on Human Factors in Electronics* (September 1962).
4. G. G. N. Wright, *The Writing of Arabic Numerals*, University of London Press, 1952.

Received September 19, 1962