# THE INFLUENCE OF REPETITION OF INCORRECTLY ANSWERED ITEMS IN A TEACHING-MACHINE PROGRAM[1,2]

*by*

JAMES G. HOLLAND AND DOUGLAS PORTER

HARVARD UNIVERSITY

One feature often considered essential to an ideal teaching machine is a mechanism which requires the student to answer each item correctly at least once. The prototype of write-in teaching machines, developed by Skinner, does provide for repetition of incorrectly answered items; but economic considerations have forced many manufacturers and experimenters to omit this feature. Commercial machines presently fail to return incorrectly answered items to the student, and the cost of a machine is greatly increased when a "memory" is included so that items missed on the first trial are repeated. Instead, current machines present each item only once. The omission of this review feature is often rationalized as inconsequential because teaching-machine programs should produce no errors, and, therefore, repetition of items should be unnecessary for "good" programs. It is self-evident that should no errors be made on a program, reviewing missed items would be superfluous. Equally self-evident, however, is the need for repeating erroneously answered items in an extremely "poor" program which generates many errors. For example, one presentation of paired-associate nonsense syllables (clearly one of the most poorly programmed types of material) would not be sufficient for acquisition of a typical, relatively short, fixed list.

This study evaluates the need for repeating missed items in a program with an error rate thought to be typical of programs for college and high school involving written answers and verbal material. The Psychology Program developed by Holland and Skinner was used. This program has had the benefit of several minor revisions and one major revision aimed at reducing the error rate. Results of the major revision indicate that the error rate has been reduced to a little over 10% of the items (Holland, 1960). The following study uses this revised version.[3]

## PROCEDURE

Fourteen graduate students in an educational psychology course were required to use this program in the Skinner write-in teaching machines as part of their course work. These student subjects were divided into two groups of seven each, matched with respect to their grade average in undergraduate study. One group used the entire program in the usual fashion, repeating missed items at the end of each item set until each item was answered correctly. The second group did not use this review feature but answered each item only once, whether or not it was missed on the first trial. Data on learning outcomes were from 3 tests which were given on appropriate sections of the teaching-machine program and were administered in the classroom. The tests consisted of completion items taken directly out of the program, selected to systematically sample both the subject-matter coverage and level of difficulty of the program. An item error count from previous teaching use of the program provided the basis for choosing test items of different levels of difficulty. Items were chosen at four levels of difficulty: The lowest level had items missed by from 0 to 20% of the students of the previous year; the second level, items missed by 21 to 40% of the students; the third level, 41 to 60%; and the fourth level, items missed by over 60% of the students. Since the program had been designed to keep errors at a minimum, it was not possible to have equal numbers of items at all four levels. Because items having a large percentage of error were so rare, only 18 of the total 196 items were available for use in the most difficult category. The others were: 36 items for the 41-60% level; 68 items for the 21 to 40% error level; and 74 items for the 0 to 20% level.

In addition to the three tests at appropriate points in the program, the same tests were given again, 6 months later, in a final examination. This unannounced retest provides data on the retention of the material and, in addition, information on possible differential forgetting by the two groups.

## RESULTS

The results of the first testing are shown in Fig. 1. The horizontal axis shows the level of difficulty of test items based on the percentage of error which occurred during previous use of the program. The vertical axis shows percentage of error on the hour tests. The two groups of *Ss*, review and nonreview, are plotted separately. At all levels of difficulty, the nonreview group missed more items than did the review group. At the lowest level of difficulty, the difference in the percentage of error is only approximately 4%; at the second level of difficulty, it is about 9%; at the third, about 12% difference; and at the fourth, the

[1]This research was reported at the American Psychological Association meeting in 1960 in Chicago.

[2]The Office of Education supported this research financially as a part of Grant No. O. E. 719067.

[3]Since completion of this study, the program (Holland & Skinner, 1961) has undergone still another major revision designed, in part, to further lower the error rate.

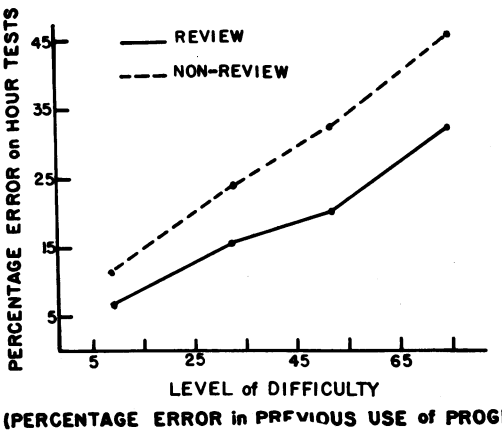


Fig. 1. The average percentage error as a function of difficulty level on the 3 tests administered immediately after the deadlines for completing appropriate sections of the program. Difficulty level is defined as the error rate on these criterion items when presented within the program during a previous use of the material.

difference is about 13%. The difference between these two groups is significant below the 1% level of confidence.[4] In addition, there is a slight trend toward somewhat greater differences between the groups for the higher levels of difficulty. Although this trend is highly suggestive and very reasonable, this interaction effect fails to reach the usual levels of statistical significance.

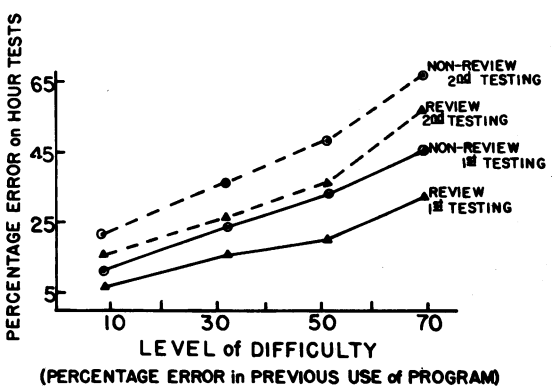


Fig. 2. The average percentage error as a function of difficulty level. The initial test and the retest after a 6-month period are shown separately for each group.

Figure 2 shows results of the retest given 6 months after completion of the teaching-machine work. Again, the level of difficulty is shown on the horizontal axis and percentage of error on the tests on the vertical axis. The two solid lines represent the same data shown in Fig. 1; the two broken lines represent the retest data for the two groups. On retest, the review group is still superior to the nonreview group. Furthermore, both groups have shown some memory decrement at all four levels of difficulty (significant at the .005 level on Wilcoxon matched-pairs test). This memory decrement is better shown by the differences in percentage of errors between the first and second testings. These differences are plotted in Fig. 3. There

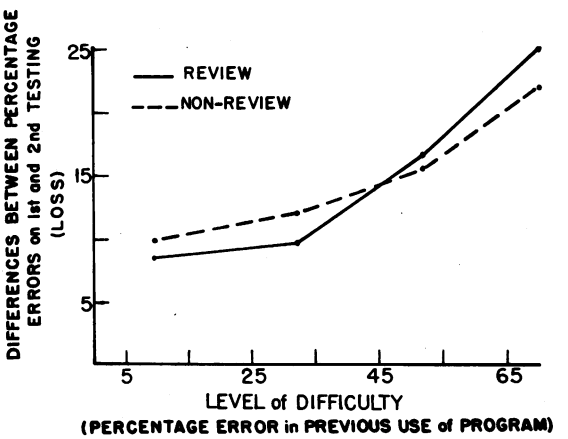


Fig. 3. The average differences in percentage errors between test and retest as a function of item difficulty.

is a loss of about 10 percentage points at the two lower levels of difficulty; and as difficulty level increases, the differences in percentage error become greater until at the most difficult level there is a loss of between 20 and 25 percentage points from the first to second testing (significant at the .01 level on the Friedman test).

## DISCUSSION

It should be remembered that the levels of difficulty represent the percentage of errors in actual use of the program; the more difficult the teaching items, the greater the memory loss. If the program had been designed so that *Ss* were ready for each item in its turn and thereby got each one correct, all items would have been at the low difficulty level (0-20% error), and the loss from the first to the second testing would have been limited to 10% or less. These data do not speak well for the "easy-come-easy-go" theory of retention, a hangover from the doctrine of "mental discipline," which has been held by some critics of teaching machines who believe that anything easily learned must be quickly forgotten. On the contrary, the present results indicate that retention is greatest when the program produces a low student error rate.

One final point to notice in the data of Fig. 3 is the almost identical loss from the first to second testings of the review and nonreview groups. Although the nonreview group performs more poorly than the re-

[4]Wilcoxon matched-pairs test, using as pairs the subscores for the matched subjects on each of the three tests and at each difficulty level.

view group on both the initial test and on the retest, there is no difference between the groups in the amount of retention decrement.

Although the review condition provides better comprehension of the material, it does require somewhat more time due to the repeating of incorrectly answered items. Table 1 shows the average time per 29-item set for the two groups. The first cycle time is the same for the two groups, but an additional 1.3 minutes per set is needed by the review group to complete the repetition of the erroneously answered items. This results in a total of 1.3 hours over the full 60 sets of items.

Table 1
Average Time in Minutes Per Set of 29 Items

| | Review Group | Nonreview Group |
|---|---|---|
| First Cycle | 12.3 | 12.3 |
| Total Cycle | 13.6 | — |

These results came as a surprise to us. We had expected very little difference between review and nonreview conditions because the error level in this program is relatively low. Furthermore, in reviewing the few missed items in each set, students often stated that they remembered the answer provided without having to read the item on its return, so that any advantage provided by the review feature would be canceled out. Apparently, however, either the memory for missed items is not so prevalent as we expected on the basis of student reports, or else short-term rote memorization of an answer is insufficient to ensure that it will be emitted again under appropriate stimulus conditions as provided by the original teaching-machine item. In the near future, at least, economics will probably prevail, and we can but regret loss of the review feature in commercial machines becoming available. The question then becomes: What can one do to overcome this deficiency; or, how can one use the additional 1.3 minutes required for review to advantage for a nonreview group? One possibility is to use diagnostic tests which will indicate review of whole lessons in areas where the student is weak, or, for a more ambitious programmer, to use parallel blocks of programmed material. A second recommendation is that of Skinner, who suggested a double stage of confirmation for each item. After the student writes his answer, the machine would reveal an additional hint for use if needed; then, if necessary, the student would write a new answer and the machine would provide the final, complete answer. Such a double-stage confirmation might make it possible to ensure that the student answers the item correctly 100% of the time.

Despite all such techniques to compensate for the elimination of the review feature, the best solution is better programs. If a program were so well written that errors seldom occurred, not only would retention be great but the review feature would be superfluous.

## REFERENCES

Holland, J. G. Teaching machines: an application of principles from the laboratory. *J. exp. Anal. Behav.*, 1960, **3**, 275-287.

Holland, J. G., and Skinner, B. F. *The analysis of behavior: a program for self-instruction.* New York: McGraw Hill Co., 1961.

*Received February 9, 1961.*