EYE MOVEMENTS AS A FUNCTION OF RESPONSE CONTINGENCIES MEASURED BY BLACKOUT TECHNIQUE¹

JUDITH DORAN AND JAMES G. HOLLAND

UNIVERSITY OF PITTSBURGH

A program may have a low error rate but, at the same time, require little of the student and teach him little. A measure to supplement error rate in evaluating a program has recently been developed. This measure, called the blackout ratio, is the percentage of material that may be deleted without increasing the error rate. In high blackout-ratio programs, obtaining a correct answer is contingent upon only a small portion of the item. The present study determined if such low response-contingent material is read less thoroughly than programmed material that is heavily response-contingent. Eye movements were compared for two versions of the same program that differed only in the choice of the omitted words. The alteration of the required responses resulted in a version with a higher blackout ratio than the original version, which had a low blackout ratio. Eighteen undergraduates received half their material from the high and half their material from the low blackout-ratio version. The order was counterbalanced. Location and duration of all eye fixations in each item were recorded by a Mackworth Eye Marker Camera. On high blackout-ratio material, subjects used fewer fixations, shorter fixation time, and shorter scanning time. High blackout-ratio material failed to evoke the students' attention.

All too familiar are the programmed instructional materials that require little of the student, teach him little, and bore him much. These are often programs with correct responses prompted by trivial cues rather than the core material that should be learned. Optimally, reinforcement (obtaining the correct answer) occurs only after completion of the precursory behavior set by the programmer. Skinner (1963) defined "programming" as "the construction of carefully arranged sequences of contingencies leading to the terminal performances which are the object of education." Acquisition of terminal performance should result from mastery of several intermediate steps increasing in difficulty as they approximate the terminal behavior. Each succeeding step in the hierarchical structure should be small enough to be within the

student's capacity at that time. A low error rate is essential because learning occurs when behavior is reinforced by the attainment of a correct answer.

However, while a low error rate is a necessary condition for material to be considered programmed, it is not a sufficient condition. A low error rate may be generated in relatively unprogrammed material by any one of a variety of prompts or trivial cues. These include repetition of the same answer for several items, providing the number of letters in the correct word and/or the first and last letters. Such trivial cues as grammar, syntax, and familiarity of phrasing may serve to "tip off" the correct answer. The following example is presented to illustrate the last type of prompt.

Representative democracy is a form of government in the United _____ of America.

One may easily obtain the correct answer without a careful reading of the entire item. A correct response reinforces whatever behavior precedes it. In some cases, this precursory behavior may be something other than the behavior prescribed by the programmer.

An additional measure is thus needed to supplement low error rate as a means of evalu-

¹This study is based on a thesis submitted by the senior author to the University of Pittsburgh in partial fulfillment of the requirement for the M.S. degree. The research was supported by the Learning Research and Development Center supported as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. Reprints may be obtained from Judith Doran, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

ating a program. The measure should be an objective, quantitative estimate of the extent to which material is actually programmed in the sense of Skinner's definition. Recently, a series of experiments developed and refined such a measure (*cf.* Holland, 1967). Applying the measure involves first judging those words and phrases that are deemed critical in obtaining a correct answer on a particular item; all material judged irrelevant in answering the item is physically obliterated by covering it with black crayon. The resulting "blackout" is considered valid if upon subsequent administration of the blacked-out program the error rate is not increased significantly. The percentage of material that can be deleted without increasing error rate is called the blackout ratio and is a measure that can be used in comparing and evaluating programs.

An item selected from each of four sets of material is shown in Fig. 1 both before and after being blacked out. The first item, from a statistics "program", is an extreme case of nearly completely unprogrammed material. The item can be answered on the basis of the expression, " $3 \times 2 \times 1 = \text{---}$," without regard to the major point of the item concerning the number of permutations of three events. Two of the items (labelled AB-A and AB-B) differ only in the word left blank; but this small difference in the design of the item makes the difference between programmed material that does not blackout and relatively unprogrammed material that blacks out substantially.

A series of experiments have demonstrated the reliability, objectivity, and validity of the blackout technique. The first of these (Holland and Kemp, 1965), used a program so replete with formal prompts that blacking out 69% of the material did not significantly increase the error rate. Two groups were given the program; one group received the original version, the other the blacked-out version. Since error rates for the two groups did not differ, the blackout was validated as a tool to distinguish programmed material from non-programmed material. To substantiate further their conclusions, Holland and Kemp prepared a "reverse blackout" version of the program, which blacked out those words and phrases previously found response contingent, *e.g.*, critical to responding correctly. Even though this reverse blackout version contained 69%

of the original material, the error rate was 72.5% compared with the 12.9% error rate on the blacked-out version containing only response-contingent material. Error rate on the original version was 11.4%.

A second study by Kemp and Holland (1966) involved a blackout analysis of 12 sets of material previously used in experiments attempting to determine whether or not there is a superiority of overt responding over covert responding. Blackouts performed on all 12 sets resulted in blackout ratios ranging from 11.1% to 74.6%. The four sets of programs favoring overt responding in previous research had the lowest blackout ratios (11.1% to 25.4%). The other eight had blackout ratios ranging from 31.0% to 74.6%. On these sets, subjects who wrote out their answers did not perform differently from subjects who did not write out their answers. Many of the sets were not programs at all because much of the material was not response contingent. Only when the critical content serves as a contingency for correct answers is the written response important.

In the programs described above, there were differences in response contingencies across programs, reflecting varying amounts of behavioral control. If the programmer wishes an effective control over the subject's behavior, the item must be constructed with care. Changing the word left out may affect error rate and reinforcement contingencies; with alteration of the required response, more or less of the material may become critical for correct responding. Holland (1964) illustrated some effects of altering the response requirement. Material from *The Analysis of Behavior* (Holland and Skinner, 1961) was altered to produce three experimental versions. The original version had a low blackout ratio and a low error rate, while one of the experimental versions had a high blackout ratio but the answers could be obtained quite easily. The second experimental version had a low blackout ratio and a high error rate. The third modified version contained no blanks and served as an "unprogrammed control". Since all four versions were identical, except for the choice of the omitted word, it was possible to determine the importance of a low blackout ratio and a low error rate by comparison on the same posttest. Only in the original version were the appropriate reinforcement contingen-

NORMAL

Stat

This illustration shows Cells 1 and 2 filled in all the possible ways they can be filled. Cell 1 = 3 ways, Cell 2 = 2 ways.

We can see in this illustration that there is only one way (ball) left to fill cell 3 when cells 1 and 2 are filled. Thus, there are $3 \times 2 \times 1 = \underline{\hspace{1cm}}$ ways in which 3 balls can fill 3 cells.

Card - V

In the word "electrocardiogram" indicate which parts mean "electrical", "heart", and "record."

AB-A

If a response which has been accidentally reinforced does not happen to occur just before a second operation of the magazine, it will be ____.

AB-B

If a ____ which has been accidentally reinforced does not happen to occur just before a second operation of the magazine, it will be extinguished.

BLACKED OUT

In the word "electrocardiogram" indicate which parts mean "electrical", "heart", and "record."

If a response which has been accidentally reinforced does not happen to occur just before a second operation of the magazine, it will be ____.

If a ____ which has been accidentally reinforced

Fig. 1. Samples of normal and blacked-out items from four different programs.

cies in effect. Errors on the posttest were significantly lower for this original version than for any of the three experimental versions. Increasing either blackout ratio or error rate resulted in poorer posttest performance. The posttest differences offer indirect evidence that subjects who used the experimental versions did not undergo the appropriate precursory behavior.

The present study was designed to determine whether material low in response contingency (thus yielding a high blackout ratio) is read less thoroughly than material high in response contingency (thus yielding a low blackout ratio). To do so, eye movements were compared for two sets of material with different blackout ratios. Actually, the two sets of material were two versions of the same

program, which varied only in the choice of the omitted word. In one version (the altered version), the blank was chosen so that the response was relatively unrelated to the rest of the item (65% blackout ratio), while in the more adequately programmed version (the original version), the response depended on more of the material (25.4% blackout ratio). If, when these programs are presented without any words covered over, the high blackout-ratio program evokes fewer eye movements, it would confirm the supposition that relatively unprogrammed material may not be read carefully. These two forms of the program had been used in demonstrations of the importance of low blackout ratio for adequate posttest performance and the importance of overt responding on a low blackout-ratio program (described above); therefore, eye movement difference in this study would tend to indicate that the difference in programmed and unprogrammed material is in the failure to engage in behavior unless it is required as precursory for the correct response.

METHOD

Subjects

Eighteen college students from Introductory Psychology classes served. All were able to read a display like those used in the experiment without glasses or contact lenses. Subjects received credit for experimental participation required of Introductory Psychology students at the University of Pittsburgh. In addition, they earned an amount of money based upon both speed and accuracy of response.

Materials

The material in the experiment proper consisted of 120 completion items, Sets 7, 8, 9, 10, from *The Analysis of Behavior* (Holland and Skinner, 1961). These sets were chosen because the necessary experimental alterations had been made on these sets for an earlier experiment (Holland, 1960, 1964) and blackout ratios had been calculated and experimentally verified (Holland and Kemp, 1965). There were two versions of Sets 9 and 10. The original version had a blackout ratio of 25.4% and the altered version a blackout ratio of 65%. In the altered version, the word omitted in the original version was completed, and another word was left blank to render the response dependent on use of less of the item. See Table 1 for examples. Each item was separately photographed and presented with a 35mm Carousel projector onto a translucent 7.5 in. (19 cm) square screen, 28 in. (71 cm) from the subject. Negatives were used giving light letters on a dark background to decrease the unpleasant glare. No item exceeded the bounds of a visual angle of 15°; each letter was approximately 3/4° visual angle in height with 1° visual angle between words.

Apparatus

A Mackworth Eye Marker Camera (Polymetric Products, Model V-1164) recorded the eye movements in the form of a motion picture of the display and a light spot reflected from the subject's left cornea. When the system is properly calibrated, the light spot is superimposed on the image of the display at a point

Table 1
Sample Items from Original and Altered Versions of Set 9 of *The Analysis of Behavior*

Original	Altered
<p><i>Item 20.</i> A school teacher is likely, when possible, to dismiss a class when her students are rowdy because she has been _____ by elimination of the stimuli arising from a rowdy class. <i>Answer Reinforced (negatively reinforced)</i></p> <p><i>Item 21.</i> The teacher who dismisses a class when it is rowdy probably causes the frequency of rowdy behavior to 1) _____ since dismissal from class is probably a(n) 2) _____ for rowdy children. <i>Answers 1) Increase 2) Reinforcement</i></p>	<p><i>Item 20.</i> A school t _____ is likely, when possible, to dismiss a class when her students are rowdy because she has been negatively reinforced by elimination of the stimuli arising from a rowdy class. <i>Answer Teacher</i></p> <p><i>Item 21.</i> The 1) _____ who dismisses a class when it is rowdy probably causes the frequency of 2) _____ y _____ r to increase, since dismissal from class is probably a reinforcement for rowdy children. <i>Answers 1) Teacher 2) Rowdy behavior</i></p>

corresponding to the point of fixation (Mackworth, 1967).

Procedure

Position of the subject in the apparatus. The subject was seated in a dental chair, his forehead rested on a brace, and his teeth were secured in a bite board covered with a dental wax impression of his teeth. His arms were folded on a ledge in front of him. Adjustments were made to obtain maximum comfort while maintaining a steady body and head position.

Calibration of the optical system. A slide containing five digits, one in the center and one in each corner, was used for calibration. The subject was instructed to fixate the center digit. The system was brought into focus, then adjusted laterally and vertically until the reflected light spot was directly upon the center digit. The subject was then asked to fixate each corner digit, after which further adjustments were made, if necessary. Recalibration on this type of slide was made after every item because the subject removed his mouth from the bite board to give his response.

Introduction of the subject to the task. The subject was next instructed in the use of his control buttons. One button opened a shutter exposing the slide and started the motor of the movie camera and another button closed the shutter and stopped the camera. The subject then pressed Button 1 and received item 1 (Set 7 of *The Analysis of Behavior*) as a sample item. Instructions were given as the subject viewed this item. As soon as he was prepared to respond, he pressed Button 2 and released the bite board. He said his answer aloud, which was recorded by the experimenter, and was then told to again press Button 1. This button press exposed a confirmation slide giving the correct answer. The next slide was a calibration slide. During the confirmation slides and the calibration slides, a small foot pedal controlled by the experimenter prevented the movie camera from operating.

Preliminary training. Each subject began with Set 7 and 8 of *The Analysis of Behavior*. For these 60 items, the subject was told that he would be paid five cents for every correct answer regardless of the time spent viewing the slide. This was done to accustom the subject to the material and the apparatus. No penalty was in effect for wrong answers at this

time. Preliminary training encompassed the first 60 items. Half of the subjects completed the 60 items in the first session. One week elapsed before the subject returned to the testing room whether or not he had finished the preliminary training. Upon his return, the subject completed any unfinished preliminary training before the experimental testing.

Experimental testing. The subjects were divided into two groups. Group 1 received the original version of Set 9 (30 items) and the altered version of Set 10 (30 items); Group 2 received the altered version of Set 9 and the original version of Set 10. This counterbalanced design aided in the control of individual differences in reading speeds and styles. Each subject was presented with a high and a low blackout-ratio program.

During experimental testing, a variable payment system encouraged speed. Table 2 presents the scale of payment for correct responding. The penalty for an incorrect response was a subtraction of three cents from the subject's earnings. The experimenter monitored a Standard Electrical clock calibrated in 1/100s of a second, which was started with the press of Button 1 and stopped with the press of Button 2. The experimenter informed the subject of how much he earned or lost after each item. Earnings ranged from \$3.45 to \$6.29 with an average of \$4.78.

Table 2
Scale of Payment for Correct Response

<i>Seconds</i> <i>Interval between Button</i> <i>Press #1 and #2</i>	<i>Cents</i>
0-5	5
6-10	4
11-15	3
16-20	2
21-	1

RESULTS

From the photographic film, data were obtained on the number and location of eye fixations per item for each subject, the average duration of fixation, and the average viewing time. Occasional calibration shifts caused incomplete data for some subjects. The subject means therefore represent different numbers of items. A chi square test performed on the total frequencies for each group was not sig-

nificant; nor was there any reason to suspect that data missing due to calibration drift was in any way differentially related to groups or conditions.

Number of Eye Movements

The major results were as expected. (See Fig. 2.) On the version with a low blackout ratio, there was a larger number of eye movements than on the corresponding altered version, which had a higher blackout ratio. The means for the two sets with low blackout ratios were 13.73 and 11.58 fixations per item (S.D. 8.95 and 9.54, respectively) as compared to the means for these two sets when the blackout ratio was high, 8.48 and 8.92 fixations per item (S.D. 9.91 and 7.21, respectively). The difference in number of fixations as a function of blackout ratio proved statistically significant ($p < 0.001$) in a repeated measures analysis of variance. In Fig. 2, there would appear to be an interaction between blackout ratio and the order in which the subjects experienced material of high and low blackout ratios. Group 1 used a set with a low blackout ratio first, then a set with a high blackout ratio,

while Group 2 had first a set with a high blackout ratio and then a set with a low blackout ratio. This interaction could result if reading patterns established on one set of material partially transferred to the next set. Thus, when the original, low blackout version of Set 10 followed the altered, high blackout version of Set 9, the frequency of fixation was lower than expected. A less-adequate reading pattern is further attested to by the high error rate (33%) for the low blackout version of Set 10, which followed the high blackout version of Set 9. (For contrast, other error rates were: 19% for the original version of Set 9, and 16% and 10% for the altered versions of Sets 9 and 10 respectively.) However, the interaction effect between order and blackout ratio fails to reach statistical significance at the 0.05 level of confidence ($F = 3.79$, F at 0.05 is 4.49). Nevertheless, the strong hint of an interaction is interesting and suggestive.

Duration of Fixation

The length of time of the single fixation was determined by counting the number of successive photographic frames with light spots in the same location. This number is a measure of time; each photographic frame represents 167 msec. The mean duration of fixation was longer on the original versions than on the altered versions. Subjects using the low blackout-ratio versions averaged 325 msec on Set 9 and 317 msec on Set 10 (S.D. 25.1 and 24.6, respectively). The same subjects given versions with a high blackout ratio averaged 290 and 295 msec (S.D. 22.4 and 22.8, respectively). This difference between high and low blackout ratios is statistically significant ($p < 0.001$, repeated measures analysis of variance). For duration of fixation, there is no hint of interaction between blackout ratio and order of presentation of units differing in blackout ratio. Subjects on the original versions fixated longer regardless of order of presentation.

Total Viewing Time

The sum per item of each photographic frame containing a light spot is a measure of the total time the subject fixated the item and is a composite of number of fixations and duration of fixation. Naturally these data also differ according to blackout ratio. The subjects fixated significantly longer periods of time on versions of low blackout ratio than on versions

Fig. 2. Mean number of eye movements on the high and the low blackout-ratio versions.

of high blackout ratio. Mean fixation per item on the two original versions were 4.52 sec and 3.62 sec (S.D. 7.77 and 8.36, respectively) as compared with 2.60 and 2.74 sec (S.D. 7.10 and 6.75, respectively) on the altered versions.

DISCUSSION

Response contingent material receives careful attention; if the subject is to answer correctly it could not be otherwise. But when less material serves as the response contingency in a high blackout program, subjects fixate less of the material and for shorter periods of time. High blackout material received less attention than low blackout material. On the program with a high blackout ratio, subjects either fixated for a short duration or not at all on the non-contingent material. The data obtained from the program with the high blackout ratio showed fewer eye movements, shorter fixation duration, and shorter scanning time per item than the data obtained from the low blackout program.

On low blackout-ratio programs, a normal reading pattern was usually found with much of the material fixated; but on high blackout programs the pattern was different. The initial fixation was often at or near the response blank and later fixations suggested a search for key words and phrases. When an occasional low blackout item was embedded in an otherwise high blackout program, the initial fixation near the blank was often followed by a large number of jumps back and forth about the item in search of cues.

The hint of a detrimental transfer from performance on high blackout material to performance on low blackout material could reflect the continuation, for a time, of these eye movement patterns. This transfer merits further investigation because, if substantiated in other experiments, it would imply that the deleterious effects of high blackout ratios reach beyond the specific material.

REFERENCES

- Holland, J. G. *Design and use of a teaching machine program*. Paper read at American Psychological Association, Chicago, September, 1960.
- Holland, J. G. Response contingencies in teaching machine programs. *Journal of Programmed Instruction*, 1964, 3, 1-8.
- Holland, J. G. A quantitative measure for programmed instruction. *American Educational Research Journal*, 1967, 4, 87-101.
- Holland, J. G. and Kemp, F. D. A measure of programming in teaching machine material. *Journal of Educational Psychology*, 1965, 56, 264-269.
- Holland, J. G. and Skinner, B. F. *The analysis of behavior*. New York: McGraw-Hill, 1961.
- Kemp, F. D. and Holland, J. G. Blackout ratio and overt responses in programmed instruction: a resolution of disparate results. *Journal of Educational Psychology*, 1966, 57, 109-114.
- Mackworth, N. H. A stand camera for line-of-sight recording. *Perception and Psychophysics*, 1967, 2, 119-127.
- Skinner, B. F. Reflections on a decade of teaching machines. *Teachers College Record*, 1963, 65, 168-177.

Received 5 November 1969.

(Revised 23 December 1970.)