EFFECTS OF REINFORCEMENT ON STANDARDIZED TEST PERFORMANCE¹

TEODORO AYLLON AND KATHY KELLY

GEORGIA STATE UNIVERSITY

The effects of two different motivational conditions upon standardized test performance were explored for two student populations. The first study involving 12 trainable retardates showed a significant increase in score on the Metropolitan Readiness Test given under reinforcement conditions when these results were compared with scores taken under standard testing conditions. In a second study, these same results were obtained with a group of 30 normal fourth-graders. An additional study was conducted to determine the effect of different experiences with token reinforcement procedures on test performance. It was found that a group of children with six weeks' exposure to reinforcement for daily academic performance scored higher under both conditions of test administration (standard and reinforcement) than a control group. However, in a single exposure to token reinforcement for correct performance on the Metropolitan Test, both the experimental group and its match control showed a parallel increase in test performance. These findings offer a procedure that yields a more representative assessment of a student's academic achievement than does testing under standard conditions.

The most widely used assessment tool in education is the standardized test. This method of sampling an individual's performance is employed on all academic levels from preschool to post-graduate and for both normal and exceptional populations. Scores from standardized tests are regressed as capsule evaluations of a child's basic academic repertoire. Kolstoe (1967) summarized the main usages of test results by contemporary educators, not the least of which is the role of test results in placement decisions.

In light of their importance, one might question whether these scores, which are treated as though they reflected the maximal academic level of the child, do, in fact, possess such per-

formance validity. Administrative procedures for individual tests, such as the Stanford-Binet, emphasize the necessity of having the optimal conditions for testing and allow, too, for specification of those variables that the test administrator recognizes as interfering with maximal effort (not enough sleep, a quarrel that day with a parent or sibling, and so on). These same considerations are not made with regard to the standardized group test score, however, even though it is recognized that the final score of any test reflects the interaction of a number of variables. Zigler (1970), for example, found a 10-point difference in I.Q. test results when the measures were taken under two different motivational conditions, yet educators maintain their heavy reliance on the standardized test score. Indeed, as early as 1925, studies were being conducted that demonstrated that praise for improvement produced greater academic performance increases than did reproof or ignoring (Hurlock, 1925).

The test administrator who has had to coax or even urge an individual to execute some

¹This paper is based in part on a thesis submitted by the junior author in partial fulfillment of the M.A. degree at Georgia State University. The authors wish to thank Mrs. Paula Calhoun and Mr. Milton White, principals of the Milton Avenue School and Jesse Mae Jones, respectively, for their cooperation and interest. Special thanks are extended to Jack Michael of Western Michigan University for his helpful suggestions in the preparation of this manuscript. Reprints may be obtained from T. Ayllon.

particular task may strongly suspect that the individual will not be performing at his optimal level. On the other hand, when a child is visibly excited and enthusiastic and expresses his wish to perform a certain activity, one may be reasonably confident that he will give his best effort when given the opportunity to engage in that activity. It would not be unreasonable, then, to hypothesize that the same behavior, carried out under each of these two conditions, would vary in either qualitative or quantitative output, or both. Although aware of these subtle but critical considerations regarding the performance of any behavior, the classroom teacher continues to administer the standardized test under antiseptic conditions as if these motivational factors did not vary. Further, the resulting data are treated as though the child had performed at his optimal level.

A situation that allowed the assessment of test performance under two or more motivational conditions would provide the basis for an empirical evaluation of motivational variables. One approach that has proved successful in establishing conditions such as these is based on principles of operant behavior that have their foundations in the experimental laboratory and which are being applied with increasing frequency to human behavior.

Following the initial application of a token economy in a psychiatric hospital (Ayllon and Azrin, 1968), this technology has been extended to a wide variety of settings. The classroom application of the token system is particularly exciting (O'Leary and Becker, 1967; O'Leary, Becker, Evans, and Saudargas, 1969; Wolf, Giles, and Hall, 1968; Birnbrauer, Bijou, Wolf, and Kidder, 1965). These reports represent only a sample of findings documenting that major changes in rates of behavior can be produced by the application of reinforcement principles to certain specified behaviors. And yet, there is little data reported that investigates the application of these procedures to the very behavior that is regarded as of major importance in educational evaluation, performance on a standardized test.

To assume that all children are motivated to perform at their peak on achievement tests would seem to be a failure to recognize the effect of at least two variables that have been shown to have major influences upon behavior, contingent reinforcement and reinforcement history. The purpose of the present study was to provide empirical data assessing the relative effect of reinforcement on test performance.

EXPERIMENT I: REINFORCEMENT TESTING WITH RETARDED CHILDREN

Twelve trainable retardates² (average I.Q. = 46.8) were given the Metropolitan Readiness Test under two test conditions. Condition I consisted of standard testing procedures as normally outlined in the administration manual. Condition II, administered on the same day, was identical, with the addition of one factor: after each subtest in the booklet, items were checked and the child received one token for each correct answer. This group had been given response and reinforcement priming to acquaint them with a token system and had exchanged their tokens for back-up reinforcers after completing the test. Test scores were raised an average of 6.25 points (of a possible 51 points), an increase significant on the 0.05 level ($t = 4.032$).

Significant increases in test scores were produced in a single exposure to reinforcement. Although these findings were statistically beyond chance occurrence, it was felt that their significance would likely be limited to the population of trainable retarded children (Zigler, 1968). Indeed, one might interpret the improvement in performance of the trainable retarded as consistent with their low level of skills and behavioral repertoire in general. Normal children, on the other hand, are likely to be func-

²A trainable retardate is defined as an individual who tests in the 30 to 55 I.Q. range and corresponds to the categories designated by the American Association on Mental Deficiency as moderately to severely retarded.

tioning at the level of behavioral complexity least affected by incentives such as those used with the trainable retarded. The purpose of Experiment II, then, was to determine the generality of the findings in Experiment I by studying a population of normal school children. Again, the major variable to be studied was test performance as it is affected by a single exposure to maximal reinforcement conditions on a standardized test. This study represents a systematic replication (Sidman, 1960) of Pilot Study I with normal children.

EXPERIMENT II: REINFORCEMENT TESTING WITH NORMAL CHILDREN

Thirty fourth-graders from an urban school were the subjects of this probe (average I.Q. = 92.8). They had taken the Metropolitan Achievement Test, Elementary Battery, which was administered under standard conditions. An alternate form of the test was also given to this class, with the addition of token reinforcement for correct responses. As with the retardates, these children receive one token for each correct answer on the test. The resulting data indicated that testing under reinforcement raised the mean performance to a statistically significant level ($t = 5.901$, $p < 0.01$).³

The data from these pilot studies indicate that the procedures using reinforcement techniques are general and consistent in their effects. Both a normal and a severely retarded population showed similar performance changes under the specified experimental conditions. The effect of

reinforcement was uniform across both groups regardless of their level of academic functioning.

In both of these studies, the subjects were introduced to their first experience with token reinforcement. Hence, the possibility exists that the dramatic effects are largely due to the novel aspects of the procedure and are not enduring over time. While the experimental procedures quickly increase performance, their continued use could result in loss of effectiveness. One way to test this argument would be actually to apply these techniques over a period of time and then to measure performance at the end of that exposure. Presumably, performance will have levelled to a natural ceiling under these motivational conditions and additional efforts to increase output would be ineffective. Thus, the use of reinforcement techniques on the standardized test itself would probably affect the test score only slightly, if at all.

Effective evaluation of this question could be achieved by using two groups, one with daily exposure to reinforcement techniques and one with standard classroom experience. At the end of a specified period of time, both groups would be tested following the procedures outlined in Pilot Studies I and II. In this way, the importance of previous experience with reinforcement techniques could be measured.

EXPERIMENT III: REINFORCEMENT HISTORY AND TEST PERFORMANCE

METHOD

Subjects

Two groups of 12 children each were matched on basis of age, I.Q., and mid-year test score on the MRT.

Setting and Staff

This study was conducted at an urban school for the trainable mentally retarded. Facilities employed were one of the regular classrooms, the school cafeteria, and the gymnasium. No special equipment was provided nor was any

³This class had also taken the Metropolitan Achievement Test at the beginning of the school year and at the end of eight months of regular school activities. Statistical analysis showed significant increases in six of the seven subtests. Nevertheless, when the same class took the Metropolitan Achievement Test under reinforcement conditions, performance scores significantly exceeded these high scoring levels in five of the seven subtests. The introduction of reinforcement paralleled the effects of eight months of classroom experience.

change made in the daily program for the rest of the classes in the school. The teacher and aide were from the regular school staff. Each performed her regular duties and no extra training was needed to implement this program. Graduate students carried out phases of the program that were conducted outside the classroom. All activities proceeded according to the standard schedule outlined for use in this school.

Experimental Design and Evaluation

The basic design utilized was the pre-program and post-program comparison of two matched groups, one under experimental treatment and the other under the standard classroom procedures for this school. The pre-program test was administered, in its entirety, under standard conditions. The post-program test was administered to the two groups in two different sessions. The first portion of the test (all odd-numbered items) was given under standard conditions. This provided for a comparison with the pre-program scores. The second portion of the post-test (all even-numbered items) was administered later the same day with maximal reinforcement for correct performance, following the procedures used in the pilot studies. These scores provided a basis for evaluation of the effects of reinforcement on test-taking performance. An analysis of variance was conducted to investigate mean differences associated with test conditions and reinforcement histories.

General Procedure

All 24 children shared the same educational program for six months. This included the following activities: academic lessons, indoor and outdoor recreation, group music activities, home-making training, and so on. At the end of this period, the Metropolitan Test was given to all children. On the basis of these scores and chronological age, the 24 children from eight classrooms were matched into 12 pairs. Twelve of these children (the experimental group) were assigned to one classroom under one teacher and were exposed to a six-week program of rein-

forcement for academic performance. The other 12 children (the control group) remained in their original classrooms and continued under the same program with no changes in procedures. Both groups took the Metropolitan Test together at the end of six weeks. Since the control group continued in their usual academic program as outlined above, there is no need for further elaboration of their activities. The experimental group, however, entered a program unique to this school. The following is a summary of their six-week experience.

Reinforcement Program with Academic Objectives

Response definition. Correct academic work selected as the target behavior for reinforcement in the experimental group included: spelling, writing, copying, arithmetic, and reading. Each day, the teacher and the aide prepared a packet of academic work for each child according to his ability level. Each packet had an equal number of pages and the children received one token per page of correct work. Tokens were delivered to the child contingent upon his correct completion of one page of academic material. The teacher and the aide delivered tokens for the written work, which was corrected as soon as the child finished it. Two opportunities were available for the child to earn tokens for academic work: during the period preceding the gym activities and also immediately before lunchtime.

Reinforcer definition. The first opportunity for token exchange took place in a special area of the gym, which had been set aside for this program. An entertainment booth was set up at which the children could purchase a variety of items and activities. The children spent 30 min each day in the entertainment area.

The second opportunity for token exchange was at lunch time. All children in the school had a hot lunch each day, which they ate in the school lunchroom. The experimental group, however, could pay admission to a special dining area adjoining the lunchroom. Admission price

was one token. The special dining area had a larger table, padded chairs, and carpeting and was cooler and quieter than the lunchroom. After lunch in the special dining area, the children could purchase cokes and candy with their tokens, as well as time to play records and watch television.

Token exchange procedure. The tokens used were painted bottle caps, which were both distinctive and inexpensive to prepare. Furthermore, they were light, easily handled, and durable, thus eliminating the need for replacement. As each child made his selection of exchange item for his tokens, he dropped the correct number of tokens into a special jar. A graduate student was in charge of the implementation of exchange procedures and also kept a record of each child's purchases per day. Colorful displays and charts gave the token prices for the various items and activities in the entertainment area.

Admission to the special dining room was also managed by the graduate student, who collected a token from each child as he entered the area and kept daily records of the purchases made by each child.

Priming. To acquaint the children with token earning and token exchange, a priming procedure was used in the settings in which each of these behaviors were to occur. In the classroom, the teacher initially explained that the children would have an opportunity to earn tokens, which could be exchanged for a variety of items that she enumerated. She then went on to describe the procedure for receiving tokens, that is, contingent upon the correct completion of each page of the work regularly given to them each day. As each token was delivered to the child, the teacher reiterated that it was for the correct work done by the child. When the children finished their work, they were instructed to take their tokens with them to the gym. In the entertainment area, the graduate students encouraged the children to play with the games and toys. This procedure follows the Reinforcer-Sampling Rule (Ayllon and Azrin, 1968), which emphasizes the importance of familiarizing the

user with a certain event in the specific context in which it is to occur. This allows the reinforcing properties of the event to be exhibited and experienced by the user. When a child continued to play with one particular item only, the graduate student told the child he could keep it (or continue playing with it, as appropriate to the particular item or activity) for a certain number of tokens. The child was then instructed as to the procedure for token exchange.

A "free day" was made available in the special dining area in order to acquaint the children with their lunchtime option. They were treated to candy, cokes, and records as well. Thereafter, the procedures for admission and for procuring the candy and cokes were followed as previously outlined.

While the experimental group was exposed to these procedures, the control group continued in their regular classroom activities. At the end of six weeks, both groups took the Metropolitan Readiness Test under the two testing conditions. Test booklets had been prepared that allowed only the odd- or the even-numbered items to be administered. The 24 children were tested on the odd-numbered items under standard conditions in one classroom. After a break for lunch, all 24 returned to the same room and received the even-numbered-item booklets. These were checked at the end of each sub-test and a token was given to the child for each correct answer.

RESULTS

Children with a six-week history of token reinforcement scored significantly higher on the portion of the test administered under standard conditions than the group in the regular academic program. A within-group analysis was made by comparing each group with its previous score. The results indicate that the experimental group averaged 3.67 points higher on the standard portion of the final evaluation than they had averaged on that same portion of the previous test. On the other hand, the control group averaged 2.75 points lower than their previous test scores. The decrease in the scores for the control

group is noteworthy because it indicates that this group, far from improving over the previous testing session, actually "lost ground" academically.

The second phase of this post-test session was conducted under reinforcement conditions. It was found that those children whose experience in school did not include token reinforcement for academic performance (control group) scored higher under reinforcement than under standard testing conditions. The average increase in score under reinforcement was 6.25 points. The group familiar with reinforcement proced-

ures showed a parallel increase in scores. The children in the experimental group averaged 28.25 on the first part under standard conditions and 35.42 on the reinforcement portion of the test, for an average increase of 7.17 points. Table 1 includes the test data for all subjects.

An analysis of variance for a two-factor experiment with repeated measures was performed on the data. Significant effects were found for both the group history variable ($F = 7.74$, $df = 1,22$, $p < 0.025$) and the testing condition variable ($F = 69.9$, $df = 1,22$, $p < 0.001$). As expected from the parallel score increases for the

Table 1
Summary of Test Scores Taken Under Two Motivational Conditions
Experiment III

S#	Pre-Program Score	Post-Program Score (Condition I)	Change Between Scores	Standard Conditions (I)	Reinforcement Conditions (II)	Change Between Scores
EXPERIMENTAL GROUP						
S-1	24	24	0	24	38	+14
S-3	22	23	+1	23	29	+6
S-5	18	20	+2	20	24	+4
S-7	24	30	+6	30	41	+11
S-9	27	31	+4	31	39	+8
S-11	30	31	+1	31	40	+9
S-13	26	30	+4	30	38	+8
S-15	22	33	+11	33	36	+3
S-17	32	39	+7	39	41	+2
S-19	31	26	-5	26	40	+14
S-21	24	30	+6	30	33	+3
S-23	14	22	+7	22	26	+4
Averages	24.58	28.25	+3.67	28.25	35.42	+7.17
CONTROL GROUP						
S-2	19	10	-9	10	24	+14
S-4	19	18	-1	18	23	+5
S-6	21	14	-7	14	21	+7
S-8	18	22	+4	22	28	+6
S-10	26	25	-1	25	34	+9
S-12	25	25	0	25	33	+8
S-14	27	21	-6	21	21	+0
S-16	27	20	-7	20	27	+7
S-18	35	34	-1	34	40	+6
S-20	29	32	+3	32	40	+8
S-22	24	23	-1	23	26	+3
S-22	25	18	-7	18	20	+2
Averages	24.58	21.83	-2.75	21.83	28.08	+6.25

two groups when reinforcement was introduced, the interaction of the two variables was not significant ($p < 1.0$). Figure 1 summarizes the results from these experimental procedures for both groups.

Fig. 1. Mean Metropolitan Readiness Test scores for each group over three test sessions (observations).

A noteworthy point regarding the effect of reinforcement on test performance is that the initial high score of the experimental group under the standard conditions might be considered as the first trial in extinction. A high score, then, is not unexpected. The introduction of reinforcement, however, served to demonstrate that even with a strong history of reinforcement, contingent reinforcement further augments test performance. The average increase for the control group was 34.6% (range: 0% to 140%). The experimental group averaged 26.5% (range: 3% to 58%).

DISCUSSION

The results from this experiment showed significantly higher Metropolitan Readiness Test scores for the token reinforcement group as compared with the children who were under standard classroom instruction. Since the comparison between the experimental and control

groups is based on performance on a standardized test, the question is what do the low scores actually reflect? Admittedly, test scores reflect a variety of conditions beyond simply "what the child knows": his health, attitude, mood, and the items themselves combine to affect the final score to a great degree. One of the objectives of this project has been to find out what happens when one of these variables is manipulated, namely, the child's motivation to perform well. Under standard conditions of testing, however, there is no means by which to determine exactly how accurate a reflection of the child's knowledge a particular test score is, for these variables cannot be eliminated or even realistically controlled. Indeed, test scores often reflect poor academic skills, but they may also reflect lack of motivation to do well in the criterion test. Conceivably, some children in this experiment learned a good deal in the six weeks of standard classroom experience (control group). Their scores taken under standard testing conditions, however, do not reflect any improvement. Only when a motivational factor was added did scores increase. Then, the academic history of the individual contributes an additional significant variable that further augments the standardized test score. These results, obtained from both a population typically limited in skills and ability as well as from a group of normal children (Experiment II), demonstrate that the use of reinforcement procedures applied to a behavior that is tacitly regarded as "at its peak" can significantly alter the level of performance of that behavior. In other words, the "typical" level of performance for that task may be lower, to a statistically significant degree, than the same performance under conditions where reinforcement procedures are applied. The considerations necessitated by these data for those who frequently use test results are clear: either the performance of the child in a standardized test situation must be maximally enhanced or the resulting test score must not be assumed to be a representative sample of the child's academic performance.

REFERENCES

- Ayllon, T. and Azrin, N. H. *The token economy: a motivational system for therapy and rehabilitation*. New York: Appleton-Century-Crofts, 1968.
- Birnbrauer, J. S., Bijou, S. W., Wolf, M. M., and Kidder, J. D. Programmed instructions in the classroom. In L. P. Ullman and L. Krasner (Eds.) *Case studies in behavior modification*. New York: Holt, Rhinehart, and Winston, Inc., 1965. Pp. 358-363.
- Hurlock, E. B. An evaluation of certain incentives used in school work. *Journal of Educational Psychology*, 1925, **16**, 145-159.
- Kolstoe, R. H. Use of test results. *Childhood Education*, 1967, **44**, 165-167.
- O'Leary, D. K. and Becker, W. C. Behavior modification of an adjustment class: A token reinforcement program. *Exceptional Children*, 1967, **33**, 637-642.
- O'Leary, D. K., Becker, W. C., Evans, M. B., and Saudargas, R. A. A token reinforcement program in a public school: a replication and systematic analysis. *Journal of Applied Behavior Analysis*, 1969, **2**, 3-14.
- Sidman, M. *Tactics of scientific research*. New York: Basic Books, 1960.
- Wolf, M. M., Giles, D. K., and Hall, R. V. "Experiments with token reinforcement in a remedial classroom." *Behavior Research and Therapy*, 1968, **6**, 51-64.
- Zigler, E. Outer-directedness in the problem-solving of retardates. In G. A. Jervis, (Ed.) *Expanding concepts in mental retardation: a symposium*. Springfield: Charles C Thomas, 1968. Pp. 123-127.
- Zigler, E. The environmental mystique: Training the intellect versus development of the child. *Childhood Education*, 1970, **46**, 402-412.

Received 28 December 1971.

(Revised 16 April 1972.)