JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1972 5, 293-309 NUMBER 3 (FALL 1972)

REWARD, COST, AND SELF-EVALUATION PROCEDURES FOR DISRUPTIVE ADOLESCENTS IN A PSYCHIATRIC HOSPITAL SCHOOL[1]

KENNETH F. KAUFMAN AND K. DANIEL O'LEARY[2]

STATE UNIVERSITY OF NEW YORK AT STONY BROOK

Sixteen pupils in a psychiatric hospital were assigned to two tutorial reading classes and balanced on six pupil characteristics and teacher preferences for the children. The effects of reward and cost procedures in a token program were assessed using both within- and between-subject comparisons in the following phases: (1) Baseline; (2) Token I, teacher evaluated and reinforced children for appropriate behavior; (3) Withdrawal of Tokens; (4) Token II, same as Token I; (5) Token III, same as Token I and II, but switched order of class meeting time; and (6) Self-Evaluation, students rated their own behavior and received prizes based on their rating, rather than the teacher's rating. The token program was markedly successful in reducing disruptive behavior and in increasing reading skills in both the Reward and Cost Classes, but there were no significant differences in the effects of the reward *versus* the cost procedure. While cost may be seen as a punishment procedure, there were no adverse side effects observed in the Cost Class at any time when the token program was in effect. The order of the classes was unrelated to the level of disruptive behavior or academic progress. The Self-Evaluative Phase, in which the students rated their own behavior, was included as an alternative to the abrupt withdrawal of tokens. In this phase, disruptive behavior remained at the previous low level.

Numerous reviews (*e.g.,* Krasner, 1971; O'Leary and Drabman, 1971; Paul, 1969) have cited the token economy as one of the more successful innovations made by behavior modifiers. In the earliest models of the token economy (*e.g.,* Ayllon and Azrin, 1968), attention was given almost exclusively to instances of acceptable behavior, while unacceptable behavior was to be ignored. Recently, however, a drastic change has taken place. A number of investigators executed token programs in which not only have tokens been awarded contingent upon desirable behavior, but they have also been taken away contingent upon the display of undesirable behavior (Burchard, 1967; Phillips, 1968; McIntire, Jensen, and Davis, *unpublished;* Boren and Coleman, 1970; Winkler, 1970). Procedures involving the subtraction of previously earned reinforcers contingent upon particular behavior are known as response-cost procedures.

Weiner (1962) performed a number of basic experiments involving cost procedures that have led some researchers to conclude that they function as a punishment procedure of considerable magnitude, similar in effect to intense electric shock (Azrin and Holz, 1966, p. 393; Kanfer and Phillips, 1970, p. 362). Considering this view of cost as a punishment procedure, cost procedures in token programs seemingly represent the antithesis of the intentions of the origi-

[1]Appreciation is due the 14 observers, and Miss Ruth Kass, who assisted in their training. The authors are also grateful for the support of Dr. Olga Von Tauber, Director, Dr. Sidney Robins, Chief Psychologist, and Mr. John Keating, Education Supervisor, of Northeast Nassau Psychiatric Hospital, Kings Park, New York. This study was adapted from a doctoral dissertation and the authors thank Leonard Krasner, Alan Ross, and Robert LeKachman for their aid. Most importantly, thanks are due Mrs. Kathryn Farrell who devoted unusual amounts of time and energy to the careful execution of the experimental procedures. Her presence during snowstorms and illness and her tolerance of verbal abuse from the pupils were especially appreciated.

[2]Reprints may be obtained from K. D. O'Leary, Psychology Department, State University of New York at Stony Brook, Stony Brook, New York, 11790; K. F. Kaufman is now at Sagamore Children's Center, Box 755, Melville, New York.

nators of token reinforcement programs who strongly avoided any punitive measures (Ayllon and Azrin, 1968). Because of the possible side effects of cost procedures it is important that research on cost procedures in applied settings be undertaken.

The purpose of the present study was to assess whether there were differential effects on academic and social behavior resulting from the application of reward and cost procedures in classroom token programs. In addition, an attempt was made to determine whether there were any detrimental side effects resulting from response-cost procedures, such as social disruption, increased aggression, or increased escape responding, as would be predicted by cost's effect as a punishment procedure (Azrin and Holz, 1966). Finally, the pupils were asked to utilize a self-evaluation procedure in order to assess whether such evaluations would be effective in maintaining desired academic and social behavior.

METHOD

Design

Adolescents were placed into one of two special reading tutorial classes, a Reward Class and a Cost Class, which met for 45 min each day, four days a week for approximately three and a half months. The classes were equivalent in as many respects as possible: *i.e.,* they used the same reading materials, they had the same teacher, and were balanced on a number of pupil characteristics. Both classes had token reinforcement programs at specified but similar times during the study. The phases of the study were as follows: (1) Baseline; (2) Token I, teacher evaluated and reinforced children for appropriate behavior; (3) Withdrawal of Tokens; (4) Token II, same as Token I; (5) Token III, same as Token I and II, but switched order of class meeting time; and (6) Self-Evaluation, students rated their own behavior and received prizes based on their rating, rather than on the teacher's rating of them.

Subjects

Sixteen pupils were chosen from the school of the adolescent unit of a psychiatric hospital who had reading deficiencies and high rates of disruptive classroom behavior. The pupils were placed into two classes balanced on the following characteristics: age, sex, psychiatric diagnosis, IQ (WISC), reading grade (Wide Range Achievement Test), and a preliminary level of disruptive behavior. Because the teacher knew many of the pupils, a further variable used to balance the classes was the pupil's rank order (1 to 16) on a questionnaire given before the study to determine the teacher's preferences of the pupils.

Teacher

The teacher, Mrs. F., was certified and had 11 yr of teaching experience, four in psychiatric institution schools. She received four graduate credits in psychology and $300 for her participation in the project.

Observation

Pupils. Every pupil was observed during each class session throughout the study. Observations were made by teams of undergraduate students. A total of 14 observers participated in pupil observation. Two teams of five observers recorded pupil behaviors each day. Because of absences and unanticipated exams of the observers, four additional observers were used throughout the study occasionally to record pupil behavior. On a given day, a maximum of five pupil observers were present. Each of four pupil observers were given random assignments to record the behavior of two pupils in each class; one pupil was observed during the first half of the class, the other pupil was observed during the second half. Thus, each pupil observer recorded the behavior of four pupils each day. A fifth observer, who acted as a reliability checker, also recorded the behavior of four pupils each day. The data collected by the pupil observers were used as the dependent measure; the data of the reliability

checker were used solely for the calculation of reliabilities.

Each pupil was observed in random order for at least 15 min per class (the median length of observation was 20 min) using the behavior codes and the method described by O'Leary, Kaufman, Kass, and Drabman (1970). Observations were made on a 20-sec observe, 10-sec record basis. The nine categories of disruptive behavior were:

1. *Out-of-chair:* movement of the child from his chair when not permitted or requested by teacher. No part of the child's body is to be touching the chair.
2. *Modified out-of-chair:* movement of the child from his chair with some part of the body still touching the chair (exclude siting on feet).
3. *Touching others' property:* child comes into contact with another's property without permission to do so. Includes grabbing, rearranging, destroying the property of another, and touching the desk of another.
4. *Vocalization:* any unpermitted audible behavior emanating from the mouth.
5. *Playing:* child uses his hands to play with his own or community property so that such behavior is incompatible with learning.
6. *Orienting:* the turning or orienting response is not rated unless the child is seated and the turn must be more than 90 degrees, using the desk as a reference point.
7. *Noise:* child creating any audible noise other than vocalization without permission.
8. *Aggression:* child makes movement toward another person to come into contact with him (exclude brushing against another).
9. *Time off task:* child does not do assigned work for entire 20-sec interval. For example, child does not write or read when so assigned.

As many as nine categories of behavior could be recorded in any 20-sec interval. Only one instance of any category of disruptive behavior could be recorded in each 20-sec interval. The daily level of disruptive behavior was calculated by dividing the total number of disruptive behavior categories recorded by the total number of intervals observed.

Teacher observation. In order to ensure that there were no differences in the teacher's behavior in the Reward and Cost Classes, and that there were no significant changes in the teacher's behavior in the different phases of the study, teacher behavior was observed and monitored daily. Teacher observations were made for 30 min during each class meeting on a 20-sec observe, 10-sec record basis. Briefly, the 11 categories of teacher behavior were:

1. *Reprimand to the class:* verbal comment indicating disapproval directed to the class as a whole or to a group of children.
2. *Praise to the class:* verbal comment indicating approval or commendation directed to the class as a whole or to a group of children.
3. *Loud reprimand to individual:* verbal comment indicating disapproval to an individual, clearly audible to the other members of the class.
4. *Soft reprimand to individuals:* verbal comment indicating disapproval to an individual which is not heard or heard with difficulty by other members of the class.
5. *Loud praise to individuals:* verbal comment indicating approval or commendation delivered to an individual in a manner clearly audible to the other members of the class.
6. *Soft praise to individuals:* verbal comment indicating approval or commendation delivered to an individual which is not heard or heard with difficulty by other members of the class.
7. *Educational attention—Close:* teacher interaction with child, primarily educational in nature, while within 3 ft from the child (excludes praise and reprimand), *e.g.,* teacher answering a question, correcting a paper when standing next to child.
8. *Educational attention—Far:* teacher interaction with child, primarily educational in nature, while further than 3

ft from child (excludes praise and reprimand); *e.g.,* while standing at front of room teacher tells child to open book to page 1.

9. *Negative facial attention:* frowning, grimacing, or eyeing down a child when behavior is not accompanied by verbal reprimand.
10. *Touching child:* touching child's person, restraining child.
11. *Redirecting attention:* diverting a child from disruptive or inappropriate behavior but making no comment about the disruptive behavior.

The first author met daily with the teacher throughout the entire study immediately after class, giving her feedback on her behavior in order to maintain stability of teacher behavior across classes and the phases of the study.

Reliability of Observations

The reliability of pupil observations was calculated both for the over-all measure of disruptive behavior and for each category of disruptive behavior. The reliability of any particular category of disruptive behavior on any day was calculated by dividing the total number of agreements by the total number of agreements plus disagreements for that category of disruptive behavior. An agreement was scored if both observers recorded the same behavior within the same 20-sec interval. A disagreement was scored if one observer recorded the behavior and the other observer did not. The reliability of the over-all measure of disruptive behavior was obtained by dividing the total number of agreements of all categories of disruptive behavior by the total number of agreements plus disagreements for all classes of disruptive behavior.

An attempt was made to obtain at least one reliability check on every pupil in both classes during each phase of the study. Because of absences of pupils or observers this goal was not met in every phase for every child. However, there were 1.84 reliability checks per child per experimental phase, indicating that reliabilities were frequently obtained. A total of 89 reliability checks were made in the Reward Class and 64 in the Cost Class during the study. The mean average reliabilities of each of the nine classes of disruptive behavior in the Reward and Cost Classes throughout the study are presented in Table 1. In addition, the reliability for the over-all measure of disruptive behavior (all behaviors) is presented in Table 1.

Table 1
Reliability of pupil observations in reward and cost classes.

Category of Pupil Behavior	*% Reliability*	
	Reward	*Cost*
Out-of-Chair	100	67
Modified out-of-Chair	100	100
Touching other's property	69	88
Vocalization	93	87
Playing	88	80
Orienting	84	80
Noise	86	80
Aggression	92	80
Time off task	94	94
All behaviors	89	86

The reliability of teacher observations was calculated in a manner identical to the calculation of pupil observations. A reliability check on teacher observations was made in both the Reward Class and the Cost Class at least once during each phase of the study. A total of 10 reliability checks were made in the Reward Class and 10 in the Cost Class. The mean average reliabilities of each of the 11 categories of teacher behavior in the Reward and Cost Classes throughout the study are presented in Table 2.

Procedure

Introduction and adjustment. The purpose of this phase included the assessment of reading skill, the establishment of classroom procedures, the introduction of the SRA Reading Laboratory Series, the adjustment of pupils to the presence of observers in the classroom, and the testing of the adequacy of the preliminary balancing of the classes.

During this phase, the procedures to be followed in all succeeding phases of the study were

Table 2
Reliability of teacher observations in reward and cost classes.

Category of Teacher Behavior	*% Reliability*	
	Reward	*Cost*
Reprimand to the class	*failed to occur*	*failed to occur*
Praise to the class	100	100
Loud reprimand to an individual	78	92
Soft reprimand to an individual	100	100
Loud praise to an individual	85	84
Soft praise to an individual	96	94
Educational attention—Close	99	99
Educational attention—Far	95	88
Negative facial attention	60	75
Touching child	81	93
Redirecting attention	75	100

introduced and established. The class was designed to be a reading tutorial, *i.e.*, the task of the pupils was to work independently at their seats on individualized reading material. The teacher's role was to work individually with those pupils who indicated they needed help. She also specified classroom rules daily, praised appropriate behavior and ignored disruptive behavior, and utilized the SRA Reading Laboratory Series. The following classroom rules were placed on a wall chart at the front of the room: "Arrive on time for class, begin work promptly, remain quiet, raise your hand to speak, stay in your seats, work hard, and no smoking."

Phase 1: Baseline. Because there were some differences between class levels of disruptive behavior during the Introduction and Adjustment Phase, the classes were rebalanced, using the levels of disruptive behavior obtained during the Introduction and Adjustment phase. Table 3 shows the average characteristics of the pupils in the two classes as they were finally composed. There were no significant differences between the classes on any of the balanced factors. After the initial matching, each class met for 10 sessions over three weeks, during which time a base rate of disruptive behavior was obtained for each of the eight pupils in both classes. The classes functioned in the manner described earlier, *i.e.*, a reading tutorial with a basic set of classroom rules, with the teacher attempting to shape appropriate behavior by praising desired behavior and ignoring disruptive behavior.

Phase 2: Token I. Assignment of classes to the Reward or Cost procedure was made by a coin toss. The differentiation of the token procedures into Reward or Cost was achieved in the following manner: in the Reward Class, each pupil began each of three 15-min rating periods with no tokens and was told that he could earn 10 tokens depending on how well he followed the rules; in the Cost Class, each pupil was given 10 "free" tokens at the beginning of each rating period and was told that the tokens were his to keep and spend, providing he followed the rules. He was also told that he could lose up to 10 tokens if the rules were not followed.

The different token systems were specifically designed so that the total number of tokens left at the end of the class for a given child displaying a given number of disruptive behaviors

Table 3
Average Characteristics of Pupils in Reward and Cost Classes

Pupil	*Age*	*Sex*	*Psychiatric Diagnosis*	*Full Scale Wisc IQ*	*WRAT READ Grade Level*	*Level of Dis. Beh Intro & Adjust. Phase*	*Mean Ranked Teacher Preference*
Reward Class	15.5	7M 1F	6 schizophrenics 2 behavior disorders	84.1	5.0	0.86*	8.6
Cost Class	15.5	7M 1F	6 schizophrenics 2 behavior disorders	85.4	5.4	0.90*	8.4

*Mean frequency of disruptive behaviors per 20-sec interval during the Introduction & Adjustment Phase.

would be exactly the same regardless of the class to which he was assigned. For example, if Pupil X were in the Reward Class and displayed three disruptive behaviors during a rating period (*e.g.,* failure to raise his hand, playing with an eraser, and talking) the teacher might have told him at the end of the rating period that he earned seven tokens. On the other hand, if Pupil Y were in the Cost Class and displayed the identical three disruptive behaviors, the teacher might have told him that he lost three tokens.

On the first day of the token program, the pupils were taken to the store, located in a separate room in the school building, and each child was allowed a few minutes to examine the items available in the store. The back-up reinforcers varied in value from less than one cent to four dollars and the prices were set so that a pupil would have to trade-in one token for every penny's worth of items he bought, *e.g.,* any item selling at 50¢ retail cost 50 tokens. A child with a perfect rating could earn approximately thirty cents worth of goods per day (per 45-min class session). Among the items available were penny candies, candy bars, inexpensive toys, stationery, health care and cosmetic items, moderately expensive toys and games, and record albums. In addition, any pupil was allowed to order any specialty item not in the store and have it reserved for him, provided that the price was less than $4.00 and there was time left in the program for him to earn enough tokens to pay for it.

Ratings were given at the end of 15-min work segments (three per 45-min class), marked by the sounding of a kitchen timer, which served as a cue for the teacher to go to the front of the room to start the procedure. While standing at the front of the room, the teacher announced to the class how many tokens each pupil was to receive. In the Reward class, she told each pupil how many tokens he had earned and what rules he had followed. Even if there were a rule infraction and she gave him fewer than the maximum number of tokens, she would attempt to explain the rating in positive terms, *e.g.:* "Johnny, you get eight tokens, because you came to class on time, you sat in your seat, you raised your hand to speak, but you worked hard only *some* of the time. You now have eight tokens." In the Cost class, she told each pupil how many tokens he had lost and what rules he had broken to account for the loss, or if he did not lose any tokens, what rule infractions he had avoided. For example, if Johnny were in the Cost Class and had demonstrated the same behavior as in the previous illustration, the teacher would have said, "Johnny, you lose two tokens. You didn't come late to class, you didn't get out of your seat, you didn't talk out, but you didn't work hard some of the time. You now have eight tokens." Following the announced ratings, she added or subtracted the indicated number of tokens from each child's plastic cylinder, which was located on a shelf at the front of the room. Finally, she recorded the ratings in a permanent record book. It is important to note that every effort was made to ensure that identical behavior received the same number of tokens in both classes. To facilitate this, guidelines for adding or subtracting points were printed on the inside cover of the Token Rating Book and the teacher consulted the guidelines before each rating. For example, the following point allotments were suggested: remaining quiet—two tokens, remaining in seat—two tokens, raising hand to speak—two tokens, sticking to the task—four tokens. It should be emphasized that the ratings were based on teacher judgment, not the observer data, because the latter method would limit the applicability of the token program.

At the end of class, all pupils were returned to the ward. To minimize the possibility of stealing, no pupils were allowed to take their tokens with them. Instead, each token jar was sealed at the end of class and then taken to the token store where the children received the back-up reinforcers after their dinner. During this phase, the Reward Class met first, from 2:45 p.m. to 3:30 p.m., followed by the Cost Class from 3:30 p.m. to 4:15 p.m. The first token phase lasted for

nine class days which, due to Christmas vacation, spanned approximately one month.

Phase 3: Withdrawal of Tokens. To demonstrate that the Reward and Cost token procedures, along with their associated back-up reinforcers, were responsible for the observed reduction in disruptive behavior, the token procedures and the back-up reinforcers were withdrawn for a two-week period. Without warning, it was announced before class on the twenty-forth day of the program that due to an inter-hospital transfer of younger children, all token programs would be temporarily suspended for at least two weeks. The pupils were assured that no one would lose previously earned tokens due to the suspension of the programs and that these would be available when the store was re-opened after the transfers had been completed. The teacher complimented the pupils on their excellent behavior in the token program and expressed the hope that the pupils would be "smart and mature" enough to continue to behave well in the coming days. The Withdrawal phase lasted for seven class days spanning approximately two weeks, during which rules, and praise and ignore remained in effect.

Phase 4: Token II. Both the Reward and Cost procedures were reinstated in their respective classes during this phase exactly as they had been during Token I. The token store was opened on each of the first three days of this phase. Thereafter, the store was opened only after every other class. The actual delay between awarding tokens and the opportunity to spend them varied between no delay and five days (due to weekends and/or holidays). There were 10 days of classes spanning approximately three weeks during this phase.

Phase 5: Token III—switched order of class meeting time. To demonstrate that any possible differences in disruptive behavior in the Reward and Cost Classes were not due to the differences in class meeting time, or the order in which the classes met (particularly since this was an after-school program), the order was switched. During this phase, the Cost Class met first, from 2:45 p.m. to 3:30 p.m. followed by the Reward Class from 3:30 p.m. to 4:15 p.m. All other aspects of the study remained as they had been during Token II. There were six days of classes spanning approximately a two-week period during this phase.

Phase 6: Self-Evaluation. After the special reading classes had been in progress for over three months, an attempt was made to examine whether changes in behavior occurred when the responsibility for evaluation of the pupil behavior is transferred from the teacher to the pupils themselves.

Token programs typically are withdrawn in one of two ways, either by a sudden cessation of tokens and back-up reinforcers as in Phase 3 of this study, or by gradually increasing the delay between the token and back-up reinforcers while simultaneously increasing the use of praise. In the past, such procedures have usually shown only limited effectiveness in maintaining appropriate behavior. Consequently, pupils were taught to evaluate their own behavior in order that such evaluation might be used as a self-control procedure in the withdrawal of token programs.

The pupils in the Reward Class were told essentially the following on the first day of this phase by the teacher:

> I am very pleased with how well you've been behaving in class lately, but it is important to learn to judge your own actions, without anyone's help. This is a very important part of becoming a mature and responsible adult. I think we can use the Special Reading Program and the token system as a way of helping to develop your ability to evaluate your own behavior.
>
> Beginning today I am going to ask you to give yourself your own ratings. You will decide how many tokens you deserve (or lose) based upon how you behaved during the rating period. In other words, I want you to determine the ratings the way I have been doing for the past weeks. You make your judgment based upon your own observation of how you followed the rules and tell me and the class how many tokens you

deserve. Of course you will *earn* (lose) as many tokens as you say. Does everyone understand what I am talking about?

This procedure was in effect for six class days for the Reward Class and seven class days for the Cost Class, spanning approximately one and a half weeks. All aspects of the procedure remained the same as they had been during Token III. The one exception was that all pupils were told at the token store on the last day of the self-evaluation phase that the Special Reading Program was ended and therefore they should spend any tokens remaining in their account.

RESULTS

Disruptive Behavior

The daily averages of the disruptive behavior for the two classes during phases one through six of the study are shown in Figure 1. Most importantly, the level of disruptive behavior was dramatically reduced for both classes during all of the token phases when compared with non-token phases. A sharp reduction in level of disruption occurred immediately in both classes on the day that the token procedures were introduced. The two classes did not differ from one another during Token I or any other phase of the program.[3] The low levels of disruptive behavior continued with remarkable stability throughout Token I. With the onset of Withdrawal, an immediate rise in disruption was apparent. By the second day of the With-

[3]Despite the non-overlapping distributions for the reward and cost classes in Token I, an analysis of covariance based upon the mean levels of disruptive behavior of each child during the baseline and Token I phases revealed no differences between the Reward and Cost procedures ($F = 0.57$, $df = 1, 13$). The only significant differences found from analyses of variance were between the Baseline and Withdrawal Phases *versus* all other phases of the study; these differences were obtained both in the Reward and Cost Classes.

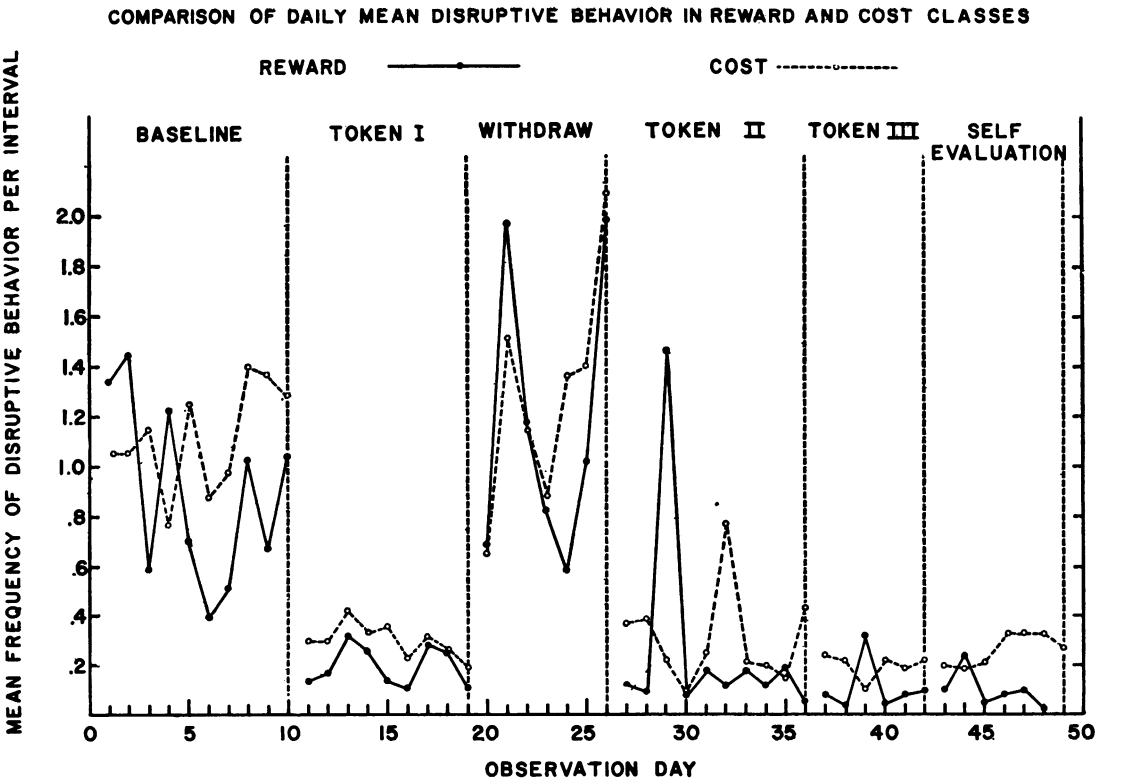

Fig. 1. Comparison of daily mean disruptive behavior in reward and cost classes.

drawal Phase, both classes had reached levels of disruptive behavior that exceeded any day during Baseline. Following this sharp rise, disruption remained at high levels throughout the rest of the phase, peaking at higher than Baseline levels once again on the final day of the phase. Although analyses of variance did not reveal a difference between the Baseline and the Withdrawal Phase for either the Reward or Cost classes, the fact that there were two days during Withdrawal on which disruption in both classes actually exceeded the Baseline was disturbing.

When tokens were restored during Token II, a marked reduction in disruptive behavior was observed. The pattern was quite similar to that found during Token I, with little variability in the levels of disruption between different days.

There were no significant differences found between Token II and Token III for either the Cost Class or the Reward Class. The level of disruptive behavior remained generally low in both classes during Token III.

During the Self-Evaluation Phase, the general low level of disruptive behavior continued in both classes. While there was no continued decrease in either the Reward or the Cost class, the maintenance of low levels of disruptive behavior during this phase was certainly striking.

An examination of the individual graphs for the Cost Class (Figure 2) reveals that the levels of disruptive behavior of seven pupils (C1, C2, C3, C4, C5, C6, C8) decreased during Cost I, increased during Withdrawal, and declined again during Cost II, and remained at a generally low level throughout the rest of the study (considering only C7's extremely low rates of disruptive behavior, he was an inappropriate subject for this study, but as will be seen later, his reading achievement (WRAT) increased dramatically). Three of the pupils (C4, C6, and C8) had their highest levels of disruptive behavior during the Withdrawal Phase.

An examination of the individual graphs for the Reward Class (Figure 3) reveals that the levels of disruptive behavior of seven of the eight pupils decreased during Reward I, while remaining essentially unchanged for one pupil (R8). Of the six pupils remaining in the study during Withdrawal Phase, five had increased levels of disruptive behavior. One subject's level of disruptiveness (R8) decreased during this phase. When tokens were reinstated during Reward II, all pupils returned to their previous low levels of disruptiveness and generally maintained these levels throughout the remainder of the study. Two pupils (R2, R3) had their highest levels of disruptive behavior during Withdrawal.

The mean percentage of occurrence of each category of disruptive behavior obtained for all pupils by class for each phase is presented in Table 4. For both classes, the mean percentage of each disruptive behavior category decreased from Baseline during Token I and returned to approximately the Baseline level during the Withdrawal Phase. A notable exception was the category "time off task", which showed a substantially higher level during Withdrawal than during Baseline in both classes.

An important aim of this study was to determine whether response cost produced detrimental side effects. One of these possible side effects might be reflected in the behavior of the pupils following a rating. If response cost were producing detrimental side effects, then the general level of disruptive behavior immediately after a rating might be greater than the level of disruptive behavior preceding the rating. In addition, greater post-rating increases in disruptive behavior would be predicted for the Cost Class compared to the Reward Class. While there was some increase in disruptive behavior following a rating, the increase was of similar magnitude for both classes. The disruptive behavior in the four intervals (approximately 2 min) after the rating was 0.5 times greater than in the four intervals preceding a token rating in both classes.

Educational Data

The effects of Reward and Cost procedures on the amount of study behavior of each of the

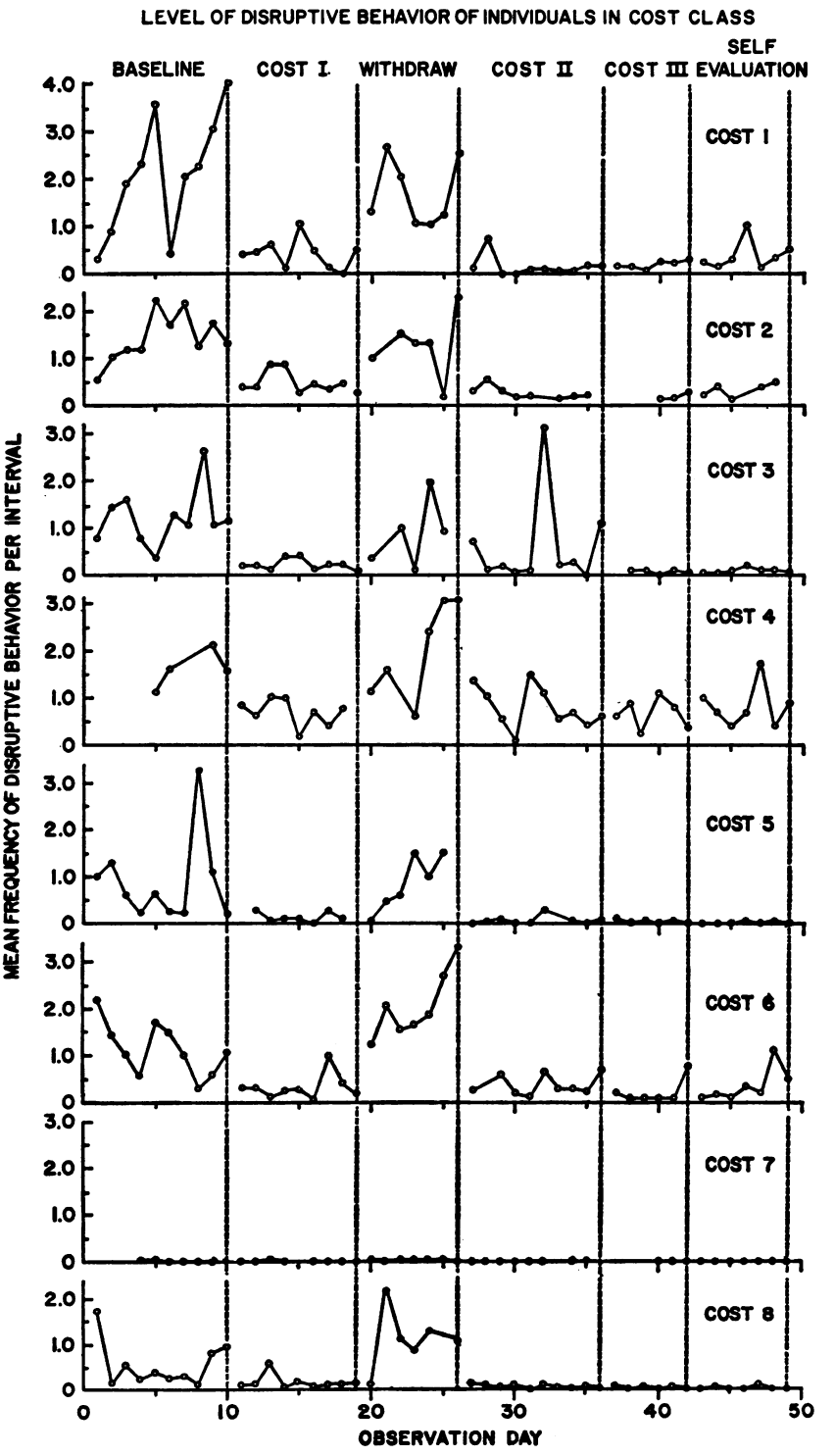

Fig. 2. Level of disruptive behavior of individuals in cost class.

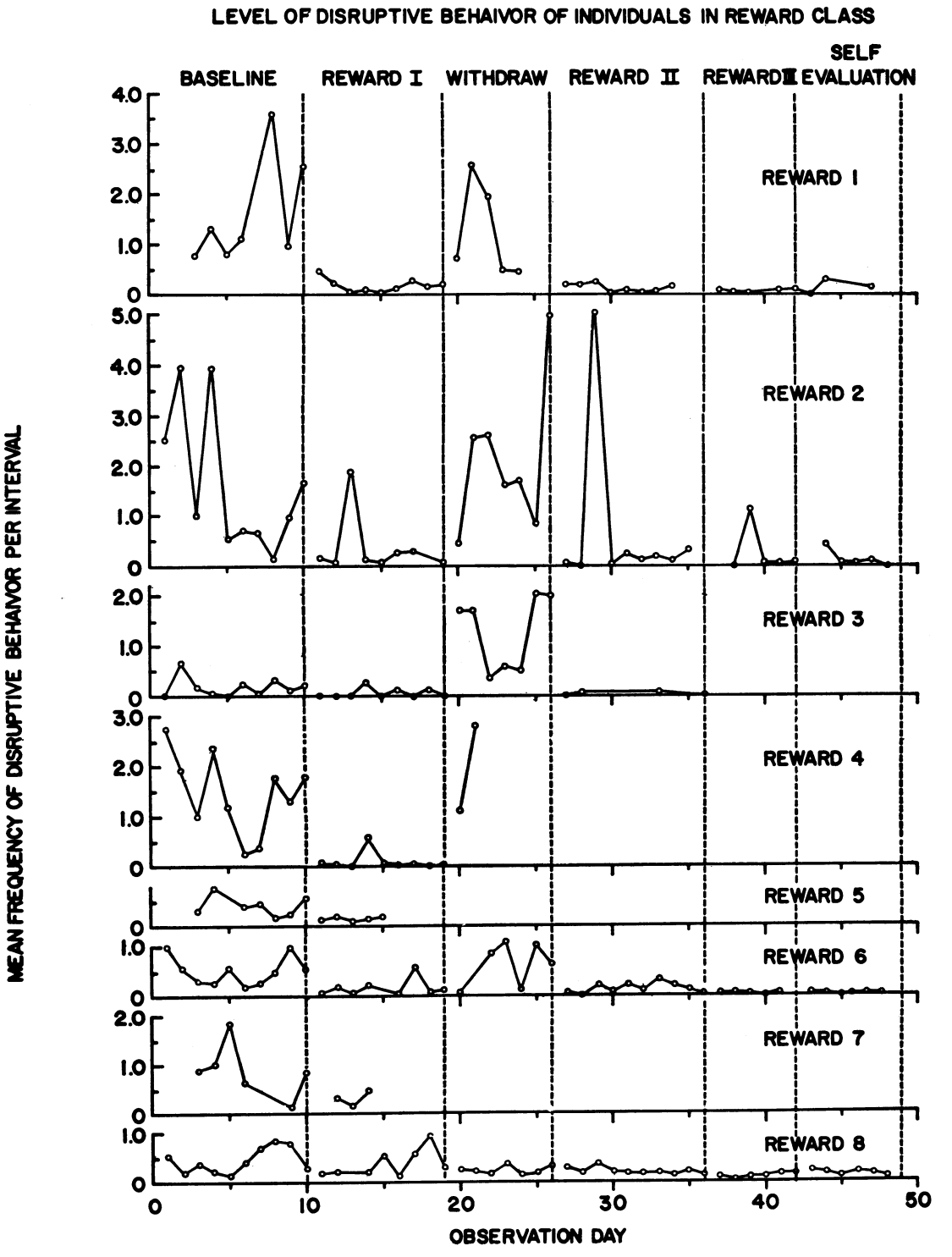

Fig. 3. Level of disruptive behavior of individuals in reward class.

Table 4

Mean percentage of occurrence per interval for each category of child behavior observed during each phase in reward and cost class.

Category	*Baseline*	*Token I*	*Withdraw*	*Token II*	*Token III*	*Self-Eval.*
	Phase					
	Reward Class					
Out of chair	3.7	0.0	2.5	0.3	0.0	0.0
Mod. out of chair	0.4	0.1	0.7	0.4	0.0	0.0
Touching	2.3	0.0	4.1	1.3	0.0	0.0
Verbalization	23.1	3.5	19.9	3.1	2.5	3.5
Playing	13.3	2.1	16.1	3.3	0.6	2.1
Orienting	13.1	3.8	14.0	6.0	1.8	1.2
Noise	16.5	3.4	11.4	3.1	1.5	0.8
Aggression	0.7	0.0	0.1	0.2	0.1	0.0
Time off task	16.7	5.5	39.6	5.7	6.5	3.0
Absence of Disruptive Behavior	54.8	82.3	48.1	85.6	89.6	90.3
All Disruptions	89.8	18.4	108.4	23.4	13.0	10.6
	Cost Class					
Out of chair	4.4	0.0	4.8	0.1	0.1	0.0
Mod. out of chair	1.4	0.1	0.1	0.0	0.0	0.0
Touching	3.6	0.1	1.4	0.6	0.0	0.2
Verbalization	29.8	5.8	28.0	6.8	3.5	5.3
Playing	15.4	5.7	17.1	4.0	2.6	3.7
Orienting	14.0	5.5	11.8	3.8	2.8	4.5
Noise	16.0	5.4	13.2	4.6	1.5	2.3
Aggression	0.9	0.1	1.8	0.1	0.0	0.1
Time off task	29.0	7.8	48.1	10.1	8.6	10.0
Absence of Disruptive Behavior	45.0	79.0	36.3	78.2	84.1	79.8
All Disruptions	114.5	30.5	126.7	30.1	19.1	26.1

pupils was evaluated by dividing the number of SRA Power Builders (Reading Selections) completed by each pupil by the total number of class sessions he attended. Although no specific consequences were made contingent upon the rapid completion of work in either class, both the Reward and Cost Classes showed overall increases in the number of Power Builders per session completed during the token phases (including Self-Evaluation Phase). Examination of individual performances revealed that four pupils in the Reward Class increased the number of Power Builders per session when token procedures were in effect, one pupil made no change, and three decreased work output during tokens (See Table 5). In the Cost Class, seven pupils increased their work output during tokens, and one decreased his output. Using a McNemar Test for the significance of changes ($N = 16$), a significant chi square was obtained (chi square $= 3.26$, $df = 1$, $p < 0.05$), indicating that more work was completed per session during the token phases than during the non-token phases of the study. However, there were no significant differences between the Reward and Cost Classes in the amount of the increase in work output during token phases ($t = 0.79$; $df = 14$).

The results of the Reading subtest of the Wide Range Achievement Test provided an additional measure of educational gain. Seven pupils in the Reward Class and six pupils in the Cost Class showed gains on the WRAT (See

Table 5
Power Builders Completed Per Session During Token and Non-Token Phases

REWARD CLASS				COST CLASS			
Pupil	*Non-Token*	*Token*	*Chge.*	*Pupil*	*Non-Token*	*Token*	*Chge.*
R1	.42	.37	−.05	C1	.29	.94	+.65
R2	.29	.55	+.26	C2	.19	.24	+.05
R3	.65	1.00	+.35	C3	.44	.95	+.51
R4	.67	1.33	+.66	C4	.00	.13	+.13
R5	.86	.40	−.46	C5	.71	1.10	+.39
R6	.63	1.17	+.54	C6	.25	1.00	+.75
R7	.67	.67	.00	C7	.64	.81	+.17
R8	.35	.13	−.22	C8	.81	.61	−.20
Mean	.57	.70	+.13	*Mean*	.42	.72	+.31

Table 6). The respective mean gains in Reading Grade Level of 0.6 for the Reward Class and 0.6 for the Cost Class are impressive because the length of the study was only 49 class sessions or the equivalent of fewer than 10 academic weeks. Even comparing these gains with the chronological length of the study (approximately 3.5 months), the gains in reading ability for the classes as a whole are impressive, and for certain pupils, it is extraordinary. The gains are even more impressive when compared with an equivalent group of seven youngsters from an adolescent unit of the hospital; they were not eligible for selection in this study due to being on different wards, and despite attending regular classes at the hospital, lost 0.2 grades on WRAT Reading during the same period of time. There were no significant differences in reading achievement gain between the respective Reward and Cost Classes ($t = 0.24$, $df = 14$).

With the increasing power of behavior modification procedures in classrooms, there is a concomitant concern about the relationship between teaching a child to behave better in a class (*i.e.*, to sit down and pay attention) and his academic output (Winett and Winkler, *in press;* O'Leary, *in press*). For example, one can cite instances of notable authors like Nabokov who purportedly did all his writing on a living room rostrum while he stood up. One wonders, is it useful to teach a child to sit still and be quiet at all? The issue is a complex one beyond the scope of this article, but some data from this study bear on the problem. Significant negative

Table 6
W.R.A.T. Reading Levels Before and After Reading Program

REWARD CLASS				COST CLASS			
	WRAT Reading Grade				*WRAT Reading Grade*		
Pupil	*Pre*	*Post*	*Diff*	*Pupil*	*Pre*	*Post*	*Diff*
R1	2.2	2.0	−0.2	C1	1.7	2.0	+0.3
R2	2.6	4.2	+1.6	C2	3.5	3.0	−0.5
R3	5.2	5.6	+0.4	C3	8.7	9.9	+1.2
R4	5.8	6.6	+0.8	C4	4.6	4.6	0
R5	5.2	5.6	+0.4	C5	6.2	6.5	+0.3
R6	6.5	7.1	+0.6	C6	6.3	7.1	+0.8
R7	7.3	7.5	+0.2	C7	5.4	7.1	+1.7
R8	5.4	6.0	+0.6	C8	6.9	8.1	+1.2
Mean	5.0	5.6	+0.6	Mean	5.4	6.0	+0.6

correlations (Spearman) were obtained between the academic products (Power Builders completed) of the children in this study and the level of their disruptive behavior. The correlations using all the children in the study across the six experimental phases were as follows: Base, $r = -0.60$, $p < 0.05$; Token I, $r = -0.50$, $p < 0.05$; Withdrawal, $r = +0.42$, NS; Token II, $r = -0.50$, $p < 0.05$; Token III, $r = -0.26$, NS; Self-Evaluation, $r = -0.52$, $p < 0.05$. In sum, generally the children who completed the most work were least disruptive.

Teacher Behavior Check

The major purpose of teacher data was to ensure relatively uniform behavior on the part of the teacher in both classes throughout the various phases of the study. The mean frequency of each category of teacher behavior per class session during phases one through six for both Reward and Cost Classes revealed little variability in teacher behavior between classes or phases (See Table 7).

DISCUSSION

The results clearly show that the Reward and Cost token procedures used in this study were both extremely effective in producing marked reductions in the disruptive behavior of hospitalized adolescents in a special reading program. In addition, both procedures led to a significant increase in educational output and reading achievement. While both the Reward and Cost procedures were highly effective, there were no significant differences found between them with respect to any of the measured variables.

The two token programs were designed to be alike in all respects, except for the method used when distributing tokens. There were only three

Table 7

Mean frequency of each category of teacher behavior per class session for reward and cost classes.

Category	*Baseline*	*Token I*	*With.*	*Token II*	*Token III*	*Self-Eval.*
			Reward			
Reprimand to class	0	0.1	0	0	0	0
Praise to class	0.8	1.1	0.6	0.6	0.7	0.7
Loud rep. to indiv.	2.8	0.1	0.4	0.2	0.2	0
Soft rep. to indiv.	0.3	0.6	0.2	0	0.2	0.2
Loud praise to indiv.	3.3	1.9	2.6	2.4	2.3	3.0
Soft praise to indiv.	3.3	3.2	2.8	2.0	2.8	2.5
Educ. atten.—Close	49.2	50.1	50.0	50.9	49.0	52.5
Educ. atten.—Far	17.9	15.8	9.6	12.6	10.5	8.3
Negative facial atten.	1.1	2.0	1.6	0.8	0.2	0.7
Touching child	4.2	1.6	4.6	1.8	3.0	0.5
Redirecting attention	4.2	0.3	2.6	0.6	0.7	0
Absence of teach. beh.	3.8	1.8	6.0	2.8	4.2	4.2
			Cost			
Reprimand to class	0.6	0.1	0	0	0	0
Praise to class	0.6	0.4	0.4	0.7	0.8	0.7
Loud rep. to indiv.	3.1	0.4	0.8	0.1	0.5	0
Soft rep. to indiv.	0.2	0.2	0.4	0	0.2	0
Loud praise to indiv.	3.7	3.1	5.0	2.8	2.3	4.5
Soft praise to indiv.	3.1	2.7	2.2	2.9	1.7	2.2
Educ. atten.—Close	52.9	50.9	54.6	50.0	52.8	50.9
Educ. atten.—Far	19.9	19.7	15.8	21.4	20.3	17.7
Negative facial atten.	1.3	1.2	2.6	1.0	0.7	0.2
Touching child	5.7	3.3	2.4	3.2	5.8	5.5
Redirection attention	5.8	0.3	4.4	0.4	3.0	1.5
Absence of teach. beh.	1.3	1.3	1.2	0.9	1.3	3.2

basic differences between the classes, each involving an aspect of the rating procedure. Briefly these were: (a) having tokens added to (Reward) or taken from (Cost) token bins, (b) being told which rules had been followed (Reward) or which rules had not been followed (Cost), and (c) being told that they had earned X number of tokens (Reward) or that they had lost X number of tokens (Cost). The failure to find differences between the effectiveness of Reward and Cost procedures may be due to the fact that the procedures were equally effective or that the procedures were not really different. The latter alternative deserves some consideration. Admittedly, it is possible that the pupils in the Cost Class ignored the previously outlined instructions and visual cues associated with the delivery and subsequent public subtraction of the "free" tokens. This is highly unlikely, if only because the procedure was repeated so many times (three times per class per pupil during the 25 days of Token I, Token II, and Token III). Furthermore, during the Self-Evaluation Phase when all pupils were required to give themselves a public verbal rating, there were no occasions when a Cost pupil failed to say, "I *lose* X tokens", or when a Reward pupil failed to say, "I *get* X tokens". One may argue, however, that since ultimately both Reward and Cost pupils obtained the same number of tokens for the identical behavior, irrespective of the class procedure, and that the amount always exceeded the number of tokens in their possession before class, then both procedures were, in reality, Reward procedures.

On the other hand, the Reward procedure may have contained certain elements that one would ordinarily associate with cost. Due to the potency of the token programs, there was very little disruptive behavior. Consequently, the teacher usually gave the maximum number of tokens during a rating (*e.g.,* 10). It is conceivable that a pupil who continually received perfect ratings came to expect these ratings in the same way as a worker paid on a salary basis expects the same pay every two weeks regardless of his work, or as a waiter who expects a 15% tip regardless of his services. A less-than-maximum rating, therefore, by its sheer infrequency, may be looked upon as a loss, *i.e.,* a cost. It is possible that a pupil receiving nine tokens in the Reward Class, may translate this rating to himself as, "Darn it, I lost one". In fact, where one has treatment programs that are equally and significantly effective, the perceptions of the recipients of the treatment may be useful to investigate.

Since the Reward and Cost Classes were at least procedurally different, and since cost procedures are considered punishment by some authors, the issue of differential side effects is important. Azrin and Holz (1966), in their comprehensive review of punishment, reject the notion that punishment procedures are ineffective. However, they state that a major undesirable feature of punishment is the social disruption that it causes. Specifically, they noted "that one side effect of the punishment process was that it reinforced tendencies on the part of the individual to escape from the punishment situation itself" (Azrin and Holz, 1966, p. 440). They predict behavior such as tardiness, truancy, dropping out of school, leaving class, increased aggression, and termination or disruption of the social relationship with the punishing agent. In the present study, an attempt was made to monitor some of the behaviors relevant to the "side effects" issue. There were no differences in the classses during the token phases on measures of achievement, aggression, or inattention as reflected in the "time off task category" or attendance (Reward 89%; Cost 91%). Because of the delay between a misbehavior and the loss of tokens in the form of a rating every 15 min, it is possible that some side effects may have been avoided. For example, one might well obtain side effects such as arguing and aggression if the loss of tokens occurred immediately after a disruptive behavior. In addition, it is possible that undesirable side effects might have occurred if the pupils could have lost earnings from amounts they had already saved, and if pupils

had lost very essential privileges such as weekend passes. In brief, while the back-up reinforcers such as records, toys, and candy were clearly desired by the pupils, it is possible that more powerful reinforcers might have generated side effects in the cost class. Boren and Coleman (1970) noted strong side effects of cost, such as rule infractions and AWOLs with soldiers who lost points for failures to attend group meetings; these men could lose already earned points for essential privileges, *e.g.,* passes. On the other hand, Phillips (1968) did not see such side effects with pre-delinquent boys who could lose similar essential privileges. Population differences, the subjects' perceived worth of the behavior that the authorities are trying to shape, and the initial dislike of the subjects toward the authorities may be of even greater importance than the aforementioned factors in producing side effects of punishment procedures.

As stated previously, increases in disruption are typically found when tokens are withdrawn. While this convenient fact is often used to verify the functional relationship between the decrease in disruption and application of token procedures (as it was in the current study), it also points out a glaring deficiency in the state of "token technology" *i.e.,* the lack of literature on fading teacher evaluation and token and back-up reinforcers and concomitant assessment of behavior. Because of the lack of research on withdrawal of token programs and the disturbingly high levels of disruptive behavior found during withdrawal, the Self-Evaluation Phase was included in the present study.

In both classes, the Self-Evaluation procedure gave the control of the token program to the pupils. They were responsible for giving themselves the token ratings that were eventually exchangeable for prizes. Essentially, this allowed the pupils the opportunity to set up a situation similar to non-contingent reinforcement. The pupils were told that they could give themselves any rating within the 10-point range. In fact, when a few pupils gave themselves less than perfect ratings, they were quickly berated as "fools" by their astonished classmates. Despite this, classroom behavior did not deteriorate.

In a certain sense, the pupils evaluated themselves correctly during self-evaluation; that is, they exhibited low rates of disruptive behavior and they generally gave themselves the highest ratings possible. On the other hand, there was no significant correlation between the pupil's evaluations and teacher's ratings. This lack of a statistically significant correlation may be the result of (1) low variability in both teacher rating and pupil evaluation and/or (2) the inability of the children to make fine discriminations about their behavior when the levels of disruptive behavior were very low.

A number of possibilities may be considered, to account for maintenance of appropriate behavior. Among them are: (a) the pupils may have found the privilege of administering their own tokens reinforcing, and therefore may have remained well-behaved in order to maintain that privilege, (b) the pupils may have suspected that failure to maintain good behavior may have led to the discontinuation of the procedure, thereby bringing about the possibility that they would get lower ratings from the teacher (or even no tokens) in the future, (c) the previous token programs may have reinforced the initiation of reading behaviors that later became reinforced by the intrinsic satisfaction associated with increasing reading skills, and, (d) since the pupils' level of disruptive behavior was low and they gave themselves high ratings to correspond with their good behavior, they have have been adventitiously reinforced for their high evaluation and good behavior.

It may very well have been that continuing the self-evaluation procedures for a longer period of time (there were six Reward Class sessions and seven Cost Class sessions) may have led to eventual increases in disruption. However, these results concerning self-evaluation provide promise for teaching self-control and for having children, rather than the teacher, assume responsibility for evaluating themselves at least during some aspects of a token program.

REFERENCES

Ayllon, T. and Azrin, N. *The token economy: A motivational system for therapy and rehabilitation.* New York: Appleton-Century-Crofts, 1968.

Azrin, N. H. and Holz, W. C. Punishment. In W. K. Honig (Ed.), *Operant behavior: areas of research and application.* New York: Appleton-Century-Crofts, 1966. Pp. 380-447.

Boren, J. J. and Coleman, A. D. Some experiments on reinforcement principles within a psychiatric ward for delinquent soldiers. *Journal of Applied Behavior Analysis,* 1970, **3,** 29-37.

Burchard, J. D. Systematic socialization: A programmed environment for the habilitation of anti-social retardates. *Psychological Record,* 1967, **17,** 461-476.

Kanfer, F. H. and Phillips, J. S. *Learning foundations of behavior therapy.* New York: Wiley, 1970.

Krasner, L. Behavior Therapy. *Annual Review of Psychology,* 1971, **22,** 483-532.

McIntire, R. W., Jensen, J., and Davis, G. *Control of disruptive classroom behavior with a token economy.* Unpublished paper presented at the meeting of the American Psychological Association Convention, San Francisco, September, 1968.

O'Leary, K. D. Behavior modification in the classroom: a rejoinder to Winett and Winkler. *Journal of Applied Behavior Analysis, (In press).*

O'Leary, K. D. and Drabman, R. Token reinforcement programs in the classroom: A review. *Psychological Bulletin,* 1971, **75,** 379-398.

O'Leary, K. D., Kaufman, K. F., Kass, R. E., and Drabman, R. S. The effects of loud and soft reprimands on the behavior of disruptive students. *Exceptional Children,* 1970, **37,** 145-155.

Paul, G. L. Chronic mental patient: Current status —future directions. *Psychological Bulletin.* 1969, **71,** 81-94.

Phillips, E. L. Achievement Place: token reinforcement procedures in a home-style rehabilitation setting for "predelinquent" boys. *Journal of Applied Behavior Analysis,* 1968, **1,** 213-223.

Weiner, H. Some effects of response cost on human operant behavior. *Journal of Experimental Analysis of Behavior,* 1962, **5,** 201-208.

Winkler, R. C. Management of chronic psychiatric patients by a token reinforcement system. *Journal of Applied Behavior Analysis,* 1970, **3,** 47-55.

Winett, R. A. and Winkler, R. C. Current behavior modification in the classroom: be still, be quiet, be docile. *Journal of Applied Behavior Analysis, (In press).*

Received 27 July 1971.
(Revised 20 November 1971.)