MULTIPLE-BASELINE ANALYSIS OF A TOKEN ECONOMY FOR PSYCHIATRIC INPATIENTS

GARY L. NELSON AND JOHN D. CONE

FULTON STATE HOSPITAL AND WEST VIRGINIA UNIVERSITY

Twelve behaviors selected for reinforcement among 16 chronic psychiatric inpatients were divided into four classes: (a) personal hygiene, (b) personal management, (c) ward work, and (d) social skills. A token economy program was introduced for each class in a sequential, cumulative, multiple-baseline format. Corrections were included for methodological deficiencies frequently encountered in past studies. Treatment variables were systematically monitored, and target behavior rates, levels of global individual functioning, general ward behavior, and off-ward behavior were assessed during baseline, implementation, and probe periods. Results indicated abrupt and substantial increases in performance of most target behaviors, significant improvements in global individual functioning ($p < .025$), positive changes in general ward behavior, and increases in social interaction during off-ward activities. The findings provide strong evidence for the efficacy of a token economy and indicate that the multiple-baseline design can be a useful method for evaluating token economy programs.

DESCRIPTORS: token economy, multiple baseline, evaluation, adult psychiatric patients

The effectiveness of token economy programs in modifying the behavior of psychiatric inpatients has been investigated in numerous studies (cf. Kazdin, 1977). Many report positive change in one or more behavioral domains, and, when considered together, the various findings suggest strongly that token economies can be an effective treatment modality. However, recent literature reviews (Gripp & Magaro, 1974; Milby, 1975; Nelson, 1977b) indicate that token economy

studies have typically suffered from important methodological deficiencies. (A notable exception is the recent report by Paul and Lentz, 1977.) In view of the shortage of methodologically adequate studies, statements regarding the effectiveness of token economy treatment should be tentative rather than conclusive.

A frequent problem with reports of token economies is technological inexactness. Procedures are not described in detail, and no data are provided to document the consistency with which they have been applied. Many investigators have summarized procedures by saying their methods were modeled after those of Ayllon and Azrin (1965). Such summaries are understandable in view of the complexity of most token economy programs. However, token programs generally represent independent attempts to apply operant principles to groups, with target behaviors, back-up reinforcers, number and type of staff, prompting techniques, and other variables differing greatly across settings. More detailed specification of the actual procedures used would provide for more meaningful interpreta-

This article is based on portions of a doctoral dissertation submitted by the first author to the Department of Psychology at West Virginia University. The authors wish to thank David Anderson for his work as project assistant, and Hugh Shallenberger and Margaret Stone for providing administrative support. We are indebted to the many individuals at Fulton State Hospital who assisted in this research. Requests for reprints should be sent to Gary L. Nelson, now at Stockton State Hospital, Stockton, California 95202. A detailed description of an expanded version of the token economy with female residents entitled "Operating a Token Economy for Long-term Psychiatric Inpatients: A Program Manual" is available through the *JSAS Catalog of Selected Documents in Psychology*, 1979, 9(1), 10. (Ms. No. 1812.)

tion of results, comparisons across programs, and refinement of treatment techniques in future work.

Another deficiency in token economy evaluations has been the rather narrow focus of their dependent measures. Assessment of changes in target behavior rates, levels of global individual functioning, and the general ward social milieu as well as measurement of generalization of change to off-ward situations and maintenance of treatment gains following discharge from the program are all important in evaluating the efficacy of token economy treatment. Most studies have fallen far short of comprehensive evaluation, assessing, at best, change in one or two of the domains mentioned above.

Perhaps the most critical methodological failing common in token economy studies has been the use of experimental designs which are not sufficiently powerful to allow demonstration of a cause-effect relationship between treatment and behavior change. Two types of designs (i.e., ABA and control group) which have been used extensively in other areas of behavior therapy research to demonstrate functional relationships have been either inappropriate or impractical in token economy evaluation. ABA designs are not appropriate when dealing with behaviors (such as grooming) which may easily come under control of nonprogrammed social reinforcers and thus not decrease substantially under withdrawal conditions. ABA designs have been used to assess changes in discrete target behaviors within a token economy but not to evaluate entire programs. Adequate control group designs are generally very impractical in applied settings because institutional considerations typically make random assignment to groups extremely difficult (Kazdin & Bootzin, 1972). An approach to token economy evaluation is needed which is sufficiently powerful to establish a functional relationship between treatment and behavior change and, at the same time, flexible enough for use by workers in applied settings.

The present study was designed to evaluate the effectiveness of a token economy through

careful attention to methodology in the areas discussed above. Four main classes of target behaviors were identified and a token economy was introduced for each class in sequential fashion to allow a multiple-baseline analysis of treatment effects. Quantitative and qualitative data regarding basic treatment variables were collected to provide an objective account of the treatment process. Measurement of changes in target behavior rates, global individual functioning, and general ward behavior was conducted. In addition, off-ward behavior was monitored to assess generalization of treatment effects.

As noted earlier, assessment of the maintenance of behavior change following discharge from a treatment program is important in assessing the full value of a treatment approach. In this study, however, our efforts were directed at providing an accurate assessment of the type of behavior change occurring during token economy treatment, a necessary step before meaningful assessment of the durability of such change can be conducted. Assessment of the maintenance of treatment gains was beyond the scope of this investigation.

METHOD

Subjects and Settings

Subjects were 16 institutionalized male psychiatric patients on a locked ward at Fulton State Hospital, Fulton, Missouri, who were present during a sufficient length of time to provide data for all periods of the study. Subjects ranged in age from 19 to 61 yr (median = 44.5 yr), and had been hospitalized from 1.2 to 41.4 yr (median = 6.9 yr). Of the subjects, 13 were diagnosed as psychotic; the other three were classified as mentally retarded. All were judged to be unsuitable for placement on an open ward, often as a result of aggressive episodes or elopement attempts. Many attended daily activity therapy sessions but otherwise received primarily custodial care. Most received psychotropic medication on a regular basis prescribed by the same physician throughout the study. Individual do-

sage levels were modified as needed, but no significant change was made in the overall use of medication during the project.

Ward staff typically consisted of three attendants on day shift, two on evenings, and two on nights. A physician, nurse, and social worker assigned to a four-ward area made ward rounds once daily. The first author (project director) and a third-year psychology graduate student (project assistant) were assigned to the ward throughout the planning and implementation of the program. In addition, three psychology staff members from other areas of the hospital functioned as observers on a part-time basis to assist in data collection.

Overview of Program Planning and Implementation

Preliminary planning. The token economy was designed to prepare residents to function on an open ward and thus move closer to release. Programming on the open ward consisted of a series of graduated steps, each involving clearly specified behavioral objectives, and a ward government system designed to promote progress through the steps by means of social recognition and peer pressure. Following informal observation and discussion with staff from both wards, target responses were selected for inclusion in the token economy that would prepare residents to perform the behavioral tasks required on the open ward and that would, therefore, have a high probability of being reinforced following transfer.

Implementation sequence. The study was carried out in four consecutive phases, each 4 wk in duration. First, a 4-wk baseline phase was conducted in which target behavior frequencies, levels of individual functioning, general ward behavior, and off-ward behavior were assessed. The target behaviors were then grouped into four categories, and a 4 wk implementation phase was initiated. During the implementation phase, the token economy was introduced to each of the four groups of target behaviors sequentially, in a multiple-baseline design, with

a new group of behaviors being added each week. Assessment of target behavior frequencies, general ward behavior, and off-ward behavior continued throughout the 4-wk implementation period.

Following the implementation phase, observation procedures not necessary to keep the program operating were discontinued for 4 wk to allow time to organize existing data and refine program procedures. During the final 4 wk (probe phase) of the project, target behavior frequencies, levels of individual functioning, general ward behavior, and off-ward behavior were again assessed while the program continued.

Target Behaviors

The 12 behaviors selected for token reinforcement were divided into four categories: personal hygiene, personal management, ward work, and social skills. The behaviors, definitions, and the number of tokens given for each occurrence following initiation of token reinforcement are presented in Table 1.

Measurement

Treatment variables (prompting, token delivery, and token store operation), target behaviors, general ward behavior, and off-ward behavior were all monitored directly by observers recording the occurrence of various categories on checklists. Individual functioning levels were assessed using a rating scale completed pre and post by ward attendants (Nurse's Observation Scale for Inpatient Evaluation, NOSIE-30, Honigfeld & Klett, 1965).

Observer training. In addition to the project director and assistant, three psychology staff members from other areas of the hospital assisted in monitoring treatment variables, target behaviors, general ward behavior, and off-ward behavior. Two of the three were kept uninformed regarding the nature of the project and functioned as blind observers to control for observer bias. Prior to data collection, all observation procedures were tested by the project direc-

Table 1
Target Behaviors, Tokens Earned, and Response Definitions

<i>Target behavior</i>	<i>Tokens</i>	<i>Response definition</i>
<i>Personal hygiene^a</i>		
Washing face ^b	1	Application of water to at least two-thirds of facial area. (Followed by drying with towel.)
Combing hair		One or more strokes through hair with a comb or brush. (Resulting in hair appearing neat.)
Shaving	1	(1) One or more strokes of razor across face or neck area. (Followed by removal of excess shaving cream and drying with a towel, resulting in clean-shaven appearance.) (2) Application of electric razor to face or neck area. (Resulting in clean-shaven appearance.)
Brushing teeth	1	(1) Insertion of toothbrush into mouth accompanied by brushing motion. (Followed by rinsing of mouth and wiping of mouth area with towel.) (2) Washing dentures in tap water.
<i>Personal Management</i>		
Dressing neatly	1	Shirt buttoned (except for top button) and tucked in; pants zipped, with belt (buckled); shoes tied, with socks. Exceptions were made for shirts or sweaters which would not normally be tucked in and for pants which required no belt.
Making bed	1 or 2	Sheet and blanket covered completely by bedspread; bedspread tucked under front of pillow then folded to headboard. Following implementation of token reinforcement, a second token was given if the bedspread was even within three inches and not touching the floor, and no obvious wrinkles were visible from 20 ft.
Cleaning bed drawer	1	All objects stacked or placed in orderly fashion; all clothes folded; no obvious dirt or dust.
Exercising	1 or 2	Physical activity of 2 or more min duration which would be likely to increase fitness (strength, endurance, or flexibility) beyond a level which would be expected from normal walking. Following implementation of token reinforcement, subjects were paid one token for performing at least half and two tokens for performing all of the exercises during a daily 10-min group exercise session.
<i>Ward Work</i>		
Examples of jobs:		Activity of 2 or more min duration which would be helpful to house-keeping or ward staff in maintaining or managing the ward. Following implementation of token reinforcement, more than 25 job descriptions were used to specify the times, locations, materials required, tasks, and supervisor review procedures for various work tasks. Pay for jobs ranged from 1 to 4 tokens.
Ashtray cleaner	1	
Cigarette roller	2	
Linen folder	4	
Dorm duster	2	
<i>Social Skills</i>		
Greeting staff	1	Initiating an appropriate verbal greeting such as "Good morning," "Hello," or "How are you?" within 30 min of the director or assistant's arrival on the ward.
Answering awareness questions correctly	3	Giving a correct answer to a preselected question regarding the immediate hospital environment or current news events (e.g., What is one advantage of moving to an open ward? Who is running for Governor?).
Verbal participation in group discussions	1 or 2	Providing information concerning specific aspects of other residents' daily behavior during a weekly ward meeting. One token was given for one instance of verbal participation, and two tokens were given for two or more instances.

^aTo facilitate observer agreement, definitions in the personal hygiene category allowed scoring of behavior with little emphasis on the quality of performance. Additional criteria to be met before receiving token reinforcement following the baseline phase appear in parentheses. The additional criteria were not included in the definitions used by observers.

^bWashing face and combing hair were paid together. Performance of both was necessary to earn one token.

tor and assistant and revised when necessary to achieve high observer agreement. The three additional observers were supplied with lists of behavioral definitions for memorization and paired with the project director or assistant during preliminary practice sessions. Typically, three practice sessions for each measure were sufficient to develop observer agreement of 85% or higher.

Two ward attendants completed the NOSIE-30 ratings. In view of the high reliability data reported for the NOSIE-30 by Kentz, Paul, and Calhoun (1971), intensive training of raters was not conducted. The project director discussed the scale with the raters and instructed each to complete the scale for several subjects before completing them for the entire group. The director then reviewed the initially completed scales and met with the raters to discuss discrepancies. Unambiguous definitions for several terms were agreed on to facilitate uniform interpretation of scale items.

Computing reliability. To obtain reliability checks for all measures other than the NOSIE-30, two observers independently observed the same persons or events simultaneously and later computed observer agreement. Percentage of agreement was computed by dividing the number of agreements by agreements plus disagreements and multiplying by 100. On measures of target behaviors and treatment variables, agreement was defined as concordance with respect to the occurrence or nonoccurrence of a particular behavior. For descriptive categories (e.g., type of prompt, prompter's manner), agreement was defined as identical endorsement of descriptors by both observers. For time-sampled data on general ward behavior and off-ward behavior, agreement was defined as concordance regarding the occurrence of a particular behavior. Agreements regarding nonoccurrence, which would tend to inflate reliability figures for these time-sampled data if included, were not scored.

To assess interrater reliability for NOSIE-30 data, product-moment correlation coefficients were computed comparing the sets of scores

produced by each rater for each of the nine NOSIE-30 subscales.

Treatment variables. Observation of treatment variables (i.e., token delivery, prompting, and token store operation) was carried out on approximately 20% of the days during the implementation and probe periods with reliability checks conducted during 20% of the observation sessions. Preceding each observation session for prompting and token delivery, specific target behaviors were selected for observation, and observers stationed themselves for likely observation of any prompts or tokens given for the selected behaviors. Observer agreement was computed separately for: (a) whether prompting or token delivery occurred, (b) the type of prompt, (c) the prompter's or deliverer's manner, (d) latency of delivery, and (e) whether tokens were offered or requested. Mean percentages of agreement ranged from 83% for the token deliverer's manner to 100% for the type of prompt given.

To monitor token store operation, observers entered the store without prior warning during a scheduled store time and marked checklists to indicate applicable descriptive categories (e.g., items available, neatness, number of customers). The mean percentage of agreement for the 13 checklist items was 90%. Data regarding treatment variables are provided in the Token Economy Program section.

Target behaviors. Response definitions in the personal hygiene category (see Table 1) emphasized enactment of specific physical operations with little regard for quality of performance. This resulted in quick, reliable determination of occurrence and allowed counting of poorly performed personal hygiene behaviors which occurred during baseline. While observers used the same definitions throughout the study, qualitative criteria were added to the definitions used for reinforcing appropriate personal hygiene behaviors with tokens.

Observers stationed themselves (usually seated) for easy observation of target behavior occurrence and assumed an unobtrusive, casual

demeanor, occasionally thumbing through magazines carried with their clipboards. They responded to residents' inquiries regarding their function with "I'm learning about the ward," and kept their interactions with residents polite but brief. Observation began at 6:40 a.m. when the residents arose, and cut-off times were designated after which specific behaviors were no longer scored (e.g., 9:00 a.m. for personal hygiene behaviors).

Most target behaviors were observed daily, Monday through Friday, except for occasional holidays, throughout the baseline, implementation, and probe phases. Data collection for answering awareness questions and cleaning bed drawers began 1 to 2 wk later than for other behaviors. Data were gathered once weekly for answering awareness questions (each subject was asked one question per wk) and twice weekly for cleaning bed drawers. Since baseline measurement of ward work and exercise involved continuous observation throughout much of the day, days were sampled at the rate of approximately two per wk during the baseline period. Ward work and exercise data were collected daily following implementation of token reinforcement because times for performance of those behaviors were then specified and observation became much simpler.

Data are not presented for verbal participation in group discussions. Because weekly progress meetings were not instituted until token reinforcement was introduced for behaviors in the social skills category, no baseline data were obtained.

Reliability checks were conducted for each target behavior during approximately 20% of the observation days. Mean percentages of agreement across days ranged from 89% (for cleaning bed drawers) to 100% (for answering awareness questions correctly) for the various target behaviors.

Global individual functioning. The NOSIE-30 was completed once during the baseline period and once approximately 3 mo later during the probe period. Correlations computed to as-

sess interrater reliability for each subscale ranged from .14 to .99 (mean = .71). The lowest correlation (.14) occurred for the Retardation subscale and was substantially lower than the correlations for all other subscales. Though the evaluation of changes in global individual functioning from NOSIE-30 ratings presented in the Results section is based on scores summed across raters, which balances out differences in the judgments of individual raters, the reliability scores are lower than expected and indicate that more extensive training of raters was needed before the study.

General ward behavior. A time-sampling procedure using the Behavior Observation Instrument (BOI) (Liberman, DeRisi, King, Eckman, & Wood, 1974) was carried out twice daily, Monday through Friday, during randomly selected time periods throughout the baseline, implementation, and probe periods. Observations were conducted only during free time periods on the ward when no structured activities were scheduled. During observation sessions, each subject was observed according to a predetermined sequence for 5 sec, and all of the behaviors observed during the 5-sec interval were checked off on the BOI rating sheet. Data on four general classes of ward behavior, each composed of several BOI categories, are presented in the Results section of this report. Reliability checks were done during approximately 10% of the observation sessions, and observer agreement was computed separately for each class of behavior. The four general classes and the BOI categories of which they were composed were: (a) social interaction (talking to resident, spoken to by resident, talking to staff, spoken to by staff); (b) constructive activity (reading or writing, individual recreation, group recreation, working, grooming or washing); (c) inappropriate behavior (repetitive motor behavior, inappropriate verbalizing, unusual body contortions); and (d) inactivity (sitting with eyes closed, lying down with eyes open, lying down with eyes closed). Mean percentages of agreement across sessions were 77% for social interaction, 95% for con-

structive activity, 71% for inappropriate behavior, and 100% for inactivity.

Off-ward behavior. Observation was accomplished using a time-sampling procedure similar to that used for assessing general ward behavior with behavioral categories altered to achieve a relevant description of off-ward activity. Behavior was observed during 1-hour recreation sessions held outdoors daily as part of an ongoing recreation program. Approximately two sessions per wk were sampled throughout the baseline, implementation, and probe periods. Reliability checks were conducted during approximately one-third of the observed sessions. Mean percentages of agreement across sessions for the four categories of behavior assessed were 72% for resident interaction, 72% for resident-staff interaction, 98% for recreational activity, and 50% for inappropriate behavior. The lowest mean agreement score (50%) reflected the low agreement scores that resulted during several sessions when only one or two instances of inappropriate behavior were observed.

A more detailed description of recording forms, observation procedures, and interrater reliability data is provided elsewhere (Nelson, 1977a).

Token Economy Program

General procedure. Before collecting baseline data, several meetings were held to orient ward staff to project goals and methods. No formal training was provided regarding behavior modification techniques. Rather, the treatment philosophy was communicated informally on a regular basis in the process of program implementation and development of ward policies, and through discussions of residents' behavior at weekly ward meetings initiated as the program progressed.

A shortage of regularly assigned attendants resulted in the ward frequently being staffed with relief personnel who were unfamiliar with token economy procedures. To ensure control and standardization of project methods for research purposes, monitoring of target behavior

performance, delivery of tokens, and operation of the token store were done by the project director and assistant and occurred between 6:40 a.m. and 3:00 p.m., Monday through Friday. (After the 4-mo research phase was completed, the program was expanded to include target behaviors reinforced by ward attendants on evenings and weekends, and responsibility for managing program operations was transferred to a BA-level psychological assistant with supervision provided by the project director.)

On Friday of the final week of baseline observation, a general ward meeting was held to describe the first stage of program implementation. The project director told subjects which behaviors were to be reinforced, when, where, by whom, and with how many tokens. To provide an introduction to tokens and their function, each subject was given enough to purchase a beverage at the token store. Immediately following the meeting, the token store was opened for the first time and beverages were made available.

On the following Monday morning, the project director or assistant was stationed immediately outside the washroom to deliver and record payment of tokens to all subjects who engaged in behaviors in the personal hygiene category. (See Table 1 for response definitions and the number of tokens given for each target behavior.) Observation of behaviors in the other three categories continued, but no tokens were given for them. Subjects were encouraged to spend their tokens at the token store.

Each Friday, a meeting was held to introduce the next category of behaviors, and token reinforcement procedures were extended to the newly added behaviors the following Monday. The order of inclusion for the four categories was: (a) personal hygiene, (b) personal management, (c) ward work, and (d) social skills.

On the Friday preceding token reinforcement for ward work, interviews were conducted with subjects to describe available chores and the number of tokens which could be earned as well as to determine work preferences. Subjects who volunteered were assigned to begin their chores

the following Monday. Chore assignments were changed monthly.

During the last week of the implementation phase, the format for the Friday ward meetings was altered to allow reinforcement of verbal participation in group discussions. The discussions which evolved were focused on the progress of individual residents and decisions regarding residents' requests for additional ground privileges were made on the basis of information provided in the discussions. Tokens were paid for verbal participation immediately following the meetings.

Reinforcers. Tokens (colored cardboard tickets, 2.5 cm by 7 cm) were delivered following occurrence of target responses. Data from observations of 124 token deliveries revealed that tokens were generally presented in a positive, congenial manner, and usually occurred almost immediately following response completion. Token delivery was accompanied by a positive statement (e.g., "Good, you did real well") 48% of the time, a neutral statement (e.g., "Here's a token for brushing your teeth") 47% of the time, and was not accompanied by any verbal statement during the remaining 5% of the observations. Within 10 sec of response completion, 46% of the deliveries were made, 32% within 10 to 60 sec, and 21% after 60 sec had elapsed. (Most instances exceeding 60 sec involved bedmaking or dressing neatly. Performance checks on these behaviors were typically conducted well after the behaviors had occurred.) Token delivery was initiated by the token deliverer during 86% of the observations and followed a reminder by the subject during 14% of the observations.

The token store was centrally located on the ward and was open during 3 15-min periods each weekday (7:10 a.m., 8:45 a.m., and 12:45 p.m.). The exchange value was between 4¢ and 5¢ per token, with subjects typically earning 9 to 12 tokens per day. Commodities in the store included hot and cold beverages, fruit, cookies, candy, ice cream, cigarettes, phonograph records, wallets, stockings, toiletries, and other miscel-

laneous items. In addition, phonograph records could be rented on an hourly basis. Systematic observation during 10 token store sessions revealed that the store was typically open on time, well stocked, and tidy. The number of customers per session ranged from 9 to 19 with a median of 15 (data regarding the number of store customers includes temporary ward residents who did not serve as subjects).

In addition to the back-up reinforcers available from the token store, shopping trips were scheduled regularly during which subjects could exchange a limited number of tokens for money to spend at local stores. Also, activities away from the hospital (e.g., the state fair, a college football game, local drama productions) were occasionally offered at a cost of 5 or 10 tokens.

Although numerous nonperishable items such as toiletries, wallets, and phonograph records were donated through the hospital Volunteer Services Department, the majority of program expenses were paid using \$25 received weekly out of hospital canteen profits.

Response cost. No token fines were included in the program design. However, restriction to the ward for 1 or more days was used as a negative consequence for several problematic behaviors including elopement, fighting, property destruction, and abuse of ground privileges. If a resident became extremely difficult to manage, he was transferred to a more restrictive ward inasmuch as there was no seclusion room available on the token economy ward.

Prompts. After token reinforcement was initiated for a category of behaviors, verbal instructions, reminders, and modeled demonstrations were frequently provided. In addition, posters were placed on the ward walls indicating target behaviors, token values, ward rules, and the ward schedule.

Observation of 130 instances of subjects performing target behaviors during the implementation and probe phases indicated that subjects received some type of individual prompt (usually a verbal reminder) by the project director, assistant, or other staff member during 24% of

the observations. The prompter's manner was always rated as positive or neutral, never negative.

RESULTS

Target Behaviors

Figures 1, 2, 3, and 4 show the daily percentages of subjects engaging in personal hygiene, personal management, ward work, and social behaviors, respectively. Since one definition covered all types of ward work, only one graph is used to display these data.

The data in Figures 1 through 4 are combined in Figure 5 to provide a multiple-baseline demonstration of the effects of the token economy on the target behaviors. Each point on the graphs represents the mean of the percentages for all of the behaviors in a category on a particular day. (In the social skills category, the weekly percentage for answering awareness questions was included in the computations for each daily overall mean during a particular week.) It may be seen that target behaviors in the personal hygiene, personal management, and ward work categories showed immediate and substantial increases following initiation of token reinforcement. These increases were maintained throughout the study. Increases for behaviors in the social skills category were less dramatic. No increases in target behavior rates were evident preceding the introduction of token reinforcement.

Global Individual Functioning

Of the 16 subjects, one was absent during pretest and two during posttest ratings of global individual functioning so that full sets of NOSIE-30 data were available for only 13 subjects. Means and standard deviations for pretest, posttest, and change scores on the NOSIE-30 subscales are presented in Table 2.

Pretest and posttest subscale scores were compared using t tests for paired observations. One-tailed t tests revealed significant increases from pretest to posttest on the following subscales:

Table 2

Means and standard deviations for subscale scores on the Nurse's Observation Scale for Inpatient Evaluation.

Subscale	Pretest	Posttest	Change
<i>Social Competence</i>			
$\bar{X}$	28.30	31.53	3.23
SD	5.17	4.86	4.28
<i>Social Interest</i>			
$\bar{X}$	19.92	24.30	4.38
SD	5.60	3.52	3.66
<i>Personal Neatness</i>			
$\bar{X}$	20.69	24.38	3.69
SD	7.29	3.92	4.19
<i>Irritability</i>			
$\bar{X}$	11.15	6.62	-4.53
SD	7.03	5.34	5.68
<i>Manifest Psychosis</i>			
$\bar{X}$	4.69	4.31	-.38
SD	3.66	4.80	4.11
<i>Retardation</i>			
$\bar{X}$	6.61	4.69	-1.92
SD	2.14	1.65	2.10
<i>Total Positive Factors</i>			
$\bar{X}$	68.93	80.23	11.30
SD	12.13	9.55	7.20
<i>Total Negative Factors</i>			
$\bar{X}$	22.46	15.62	-6.84
SD	9.64	8.71	9.66
<i>Total Patient Assets</i>			
$\bar{X}$	142.46	160.61	18.15
SD	19.25	16.15	15.33

Social Competence, $t(12) = 2.63$, $p < .025$; Social Interest, $t(12) = 4.83$, $p < .001$; Personal Neatness, $t(12) = 3.17$, $p < .005$; Total Positive Factors, $t(12) = 5.65$, $p < .001$; and Total Patient Assets, $t(12) = 4.26$, $p < .001$. Significant decreases were found for Irritability, $t(12) = 2.88$; $p < .01$, Retardation; $t(12) = 3.30$, $p < .005$; and Total Negative Factors, $t(12) = 2.55$, $p < .025$. Thus, the subjects exhibited significant changes in the direction of improved functioning on all subscales except Manifest Psychosis, where no significant change was noted.

General Ward Behavior

One hundred BOI observation sessions were conducted, 32 during baseline, 34 during implementation, and 34 during the probe period. The total percentages of observed behavior accounted

PERSONAL HYGIENE

Fig. 1. Percentages of subjects performing personal hygiene behaviors.

PERSONAL MANAGEMENT

Fig. 2. Percentages of subjects performing personal management behaviors.

WARD WORK

Fig. 3. Percentages of subjects performing ward work.

for by each of the four general classes were computed for each period and are displayed in Figure 6. Changes in all classes of behavior were in the direction of improved ward behavior, and improvement continued over the three measurement periods.

Off-Ward Behavior

Of the 16 off-ward observation sessions conducted, 4 were during baseline, 6 during implementation, and 6 during the probe period. The total percentages of observed behavior accounted for by each of the four general classes of off-ward behavior during baseline, implementation, and probe were, respectively, 9%, 14%, and 19% for resident interaction; 14%, 6%, and 8% for resident-staff interaction; 26%, 19%, and 32% for recreational activity; and 4%, 1%, and 1% for inappropriate behavior.

Changes were in the general direction of improved functioning, but were not dramatic. A clear trend toward increased interaction among subjects was evident. The increase in social interaction among residents may have been partly due to the fact that there were fewer staff to interact

with during the implementation and probe periods. However, the combined percentages for resident and resident-staff interactions increased from 23% during baseline to 27% during the probe period, indicating a general increase in interaction with others as the study progressed.

Resident Movement

Although data for this study are derived from only the 16 residents who were present during all periods of the project, a steady flow of other residents moved to and from the ward during the 4-mo period. Of these, four left during the baseline month. One of the moves was positive (i.e., to an open ward, to placement outside of the hospital), two were negative (i.e., to a more restrictive ward), and one resident eloped from the hospital. During the first 3 mo of treatment (implementation through probe phases), 17 residents left the ward. Nine of the moves were positive, six negative, two were elopements. Thus, positive transfers occurred more frequently following the baseline month, and negative transfers and elopements occurred at approximately the same rate.

SOCIAL SKILLS

Fig. 4. Percentages of subjects performing social skills behaviors.

DISCUSSION

Substantial improvements were noted in the performance of most target behaviors following implementation of the token economy program. The only target behavior which did not show a clear increase was answering awareness questions correctly. It could not be determined here whether the lack of change was due to inadequate reinforcement procedures or to the measurement procedures not being accurate or sensitive enough to reveal changes which might have occurred.

Not only did the subjects perform target behaviors at substantially greater rates, their over-

all individual functioning was judged by ward attendants to be significantly improved, as indicated by higher posttest scores on eight out of nine NOSIE-30 subscales. Also, gradual and consistent improvements in general ward behavior were noted as the study progressed. Immediate or dramatic changes in general ward behavior were not expected because observation sessions were conducted during free-time periods and did not include busy times when target behaviors were being reinforced.

Behavioral changes during off-ward activity sessions were less pronounced than the improvements evident in general behavior on the ward but did indicate a general increase in social inter-

OVERALL PROGRAM

Fig. 5. Overall mean percentages of subjects performing personal hygiene, personal management, ward work, and social skills behaviors.

action. As observed changes in off-ward behavior were neither dramatic nor totally consistent in their direction, evidence that treatment gains

noted on the ward generalized to off-ward situations is tentative.

Resident movement data indicated that trans-

Fig. 6. Percentages of observed ward behavior accounted for by social interaction, constructive activity, inappropriate behavior, and inactivity.

fers to more desirable environments increased following the baseline month. However, while transfer data are valuable, they do not provide an adequate measure of treatment effectiveness because transfers often depend on factors other than the residents' behavior (e.g., availability of space on other wards).

Baseline target behavior rates were stable for all categories until token reinforcement was introduced to each in turn. Thus, the multiple-baseline design allowed demonstration of functional relationships between token economy procedures and increases in target behavior rates. The link established between the token program and improvements in global individual functioning, general ward behavior, and off-ward behavior was correlational, however, because assessment in those three domains was not part of the

multiple-baseline analysis. Although evidence of a direct causal link between the token economy and behavior changes other than those directly reinforced was not provided, it is reasonable to hypothesize that such changes occurred as a result of the token economy program since the effect of the program on a variety of target behaviors was clearly established.

The results of this study suggest, on the one hand, that changing specific behaviors can result in a generalized improvement in other areas of behavioral functioning. On the other hand, the fact that behavior rates in each of the four categories remained stable until token reinforcement was introduced to each in sequence indicates that discrete categories of behavior within a token economy can be identified which are relatively independent of one another, at least over brief

time periods. Kazdin (1973) discussed the possibility of response generalization as a potential difficulty in the use of multiple-baseline designs. In the present study, the grouping of behaviors and implementation sequence were carefully planned to minimize response generalization or other unplanned increases in target behavior rates. Behaviors requiring the most conspicuous observation procedures were reinforced first because the staff's interest in those behaviors would be obvious once the program began. The second behavioral category to be reinforced included behaviors which could be observed in a relatively unobtrusive manner. Ward work was one of the last two categories to be reinforced because the authors' previous experience with token economy implementation indicated that work behavior would not increase substantially prior to the identification and description of specific chores to be reinforced. Token reinforcement was introduced last for behaviors in the social skills category because the authors had little experience targeting such behaviors within a token program and were unsure of the outcome.

Multiple-baseline designs have typically involved introduction of a treatment condition to three or four target behaviors, individually, in sequential fashion. To proceed one behavior at a time in the implementation of a full-scale token economy would be difficult as well as time consuming. Moreover, the probability of inter-correlations among responses would increase as a function of the number of target behaviors included and thus make the demonstration of functional relationships more difficult. By grouping related behaviors and introducing treatment to the behaviors in classes, rather than singly, the multiple-baseline design can be appropriately applied to a large-scale program in a relatively brief time period.

Based on information provided in this investigation, it is possible to estimate the costs (over and above existing costs) required to operate a similar token economy on a medium-sized psychiatric ward (20 to 35 beds). Although a variety of staffing patterns could be used, a logical

plan based on our experience would include: (a) one full-time BA-level psychological assistant to manage the daily operation of program procedures (e.g., prepare and maintain data forms, record and reinforce resident behavior, monitor recording and token reinforcement conducted by other staff, stock the token store, supervise store operation, assign ward chores, etc.) and (b) one half-time psychologist with training in behavior analysis to monitor overall program functioning, plan revisions where needed, develop individualized programming, train staff in the use of behavioral techniques, and provide supervision for the program manager (psychological assistant). The above-named personnel would be in addition to staff normally required to operate a ward without a token economy program (e.g., several ward attendants per shift, and a physician, a social worker, a nurse, and an activity therapist on a part-time basis). Costs other than salaries for the additional staff described above would include approximately \$100.00 worth of initial supplies (tokens, a coffee maker, a used refrigerator, display shelves, a service counter, etc.), and approximately \$25.00 weekly to keep the token store adequately stocked.

Though the study documents program effectiveness over a relatively brief period (3 mo), subsequent checks of ongoing token economy records 1 yr later indicated that performance of target behaviors continued at high rates. Also, only six subjects involved in the study remained; the majority had moved on to open wards or left the hospital. At the time of this writing, the token economy was well established as an important component of a larger system of coordinated ward programming designed to move residents from the hospital to community living (Stone & Nelson, in press).

REFERENCES

- Ayllon, T., & Azrin, N. H. The measurement and reinforcement of behavior of psychotics. *Journal of the Experimental Analysis of Behavior*, 1965, 8, 357-383.

- Gripp, R. F., & Magaro, P. A. The token economy program in the psychiatric hospital: A review and analysis. *Behavior Research and Therapy*, 1974, 12, 205-228.
- Honigfeld, G., & Klett, C. J. The nurses observation scale for inpatient evaluation: A new scale for measuring improvement in chronic schizophrenia. *Journal of Clinical Psychology*, 1965, 21, 65-71.
- Kazdin, A. E. Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis*, 1973, 6, 532-539.
- Kazdin, A. E. *The token economy: A review and evaluation*. New York: Plenum Press, 1977.
- Kazdin, A. E., & Bootzin, R. R. The token economy: An evaluative review. *Journal of Applied Behavior Analysis*, 1972, 5, 343-372.
- Lentz, R. J., Paul, G. L., & Calhoun, J. F. Reliability and validity of three measures of functioning with "hard-core" chronic mental patients. *Journal of Abnormal Psychology*, 1971, 78, 69-76.
- Liberman, R. P., DeRisi, W. J., King, L. W., Echman, T. A., & Wood, D. Behavioral measurement in a community mental health center. In P. O. Davidson, F. W. Clark, and L. A. Hamerlynck (Eds.), *Evaluation of behavioral programs in community, residential, and school settings*. Champaign, Ill.: Research Press, 1974.
- Milby, J. D. A review of token economy treatment programs for psychiatric inpatients. *Hospital and Community Psychiatry*, 1975, 26, 651-658.
- Nelson, G. L. Multiple-baseline analysis of a token economy for psychiatric inpatients. (Doctoral dissertation, West Virginia University, 1976). *Dissertation Abstracts International*, 1977, 37, 4158B. (University Microfilms No. 77-2571, 186). (a)
- Nelson, G. L. *Methodology in token economy evaluation: A review and critique*. Morgantown: West Virginia University, 1977. (ERIC Document Reproduction Service No. ED 147 711). (b)
- Paul, G. P., & Lentz, R. J. *Psychosocial treatment of chronic mental patients*. Cambridge, Mass.: Harvard University Press, 1977.
- Stone, M. E., & Nelson, G. L. Coordinated treatment of long-term psychiatric inpatients. *Social Work*, in press.

Received February 15, 1978

Final acceptance September 22, 1978