*THE BEHAVIOR OBSERVATION INSTRUMENT: A METHOD
OF DIRECT OBSERVATION FOR PROGRAM EVALUATION¹*

PETER ALEVIZOS, WILLIAM DERISI, ROBERT LIBERMAN,
THAD ECKMAN, AND EDWARD CALLAHAN

CAMARILLO-NEUROPSYCHIATRIC INSTITUTE (UCLA) RESEARCH CENTER,
OXNARD MENTAL HEALTH CENTER, AND CAMARILLO STATE HOSPITAL

The background and development of a multicategory direct observation system, the Behavior Observation Instrument (BOI), is described. This time-sampling procedure for recording the behavior of persons is demonstrated in several treatment settings and the results applied to issues of program evaluation. Elements that have prevented direct observation from being widely adopted, such as costs, manpower, and training requirements, are systematically analyzed. A basic psychometric analysis of the instrument is used to determine optimum frequency and duration of observation intervals as well as observer agreement. The results imply that direct observation methods, once assumed by some to belong to the special province of the single-subject design, can be used to assess the effects of programs on groups of psychiatric clients in an efficient and economic manner.

DESCRIPTORS: program evaluation, recording and measurement techniques, methodology, direct observation, time sampling, cost analyses, inpatient and day treatment settings, psychiatric patients

As political and funding priorities increasingly focus on cost-effectiveness of mental-health treatment, demands for "hard" data on program

evaluation will escalate. The Nader group report on the NIMH stated that: "In retrospect . . . NIMH made only the most piddling evaluation of the program's performance" (Chu and Trotter, 1972, p. 11). In California, local mental-health programs were required to submit evaluations of their services by 1975 to continue receiving the funds from the state government, which supports 90% of their programs (California State Legislature, 1971). Clinicians and administrators will be pressed to find both treatment and evaluation services that are cost-efficient in their use of resources per client served. Viewed in the light of economic reality, behavior analytic procedures that consume large amounts of a program's resources, or that are

¹The work in this report was supported in part by NIMH Research Grant Number MH 26207-01 from the Mental Health Services Research and Development Branch and NIMH Hospital Improvement Project Grant Number MH-R 20-C to Robert Paul Liberman, Principal Investigator. The opinions stated are those of the authors and do not reflect official policy of the State of California, Department of Health, Ventura County Health Services Agency, or Regents of the University of California. The authors are grateful for the cooperation of the clinical staffs at Camarillo State Hospital, Program 2R and at the Oxnard Mental Health Center, and for the support and encouragement of Clinton Rust (Executive Director, Camarillo State Hospital), Rafael Canton, M.D., and Sarah Miller, M.D. (respectively, past and present Director of Ventura County Mental Health Services), and Robert Coombs, Ph.D. (Chief of Research, Camarillo-NPI Research Center). Research assistants deserving special credit for conducting the behavioral observations are: Philip Berck, Michael Campbell, Aurora de la Selva, Veronica Fabian, Paul Gabrinetti, Mary Hernandez, Susan Heyl, Sandranne Lenhardt, Melinda Maggiani, Richard Schultz, and David Wood. The research reported is the result of work on two separate but co-

ordinated projects. The order of authorship was assigned alphabetically for P. Alevizos and W. DeRisi, directors of the two projects, and neither claims primacy or seniority. William DeRisi is now at the California Department of Health, and Edward Callahan is now at The University of West Virginia. Reprints may be obtained from Peter N. Alevizos, Department of Psychology, Straub Hall, University of Oregon, Eugene, Oregon 97403.

narrowly confined to a few clients or client behaviors, are likely to be discarded in favor of less expensive rating procedures.

Behavioral assessment methods are not only useful in demonstrating the effectiveness of individual treatment, but are also suited to the evaluation of large-scale treatment programs in a variety of settings. Multidimensional or multicategory behavioral measures are needed to detect both positive and unintended effects in the repertoire of individuals targeted for some specific behavior change, as well as wider positive or negative effects on the behavior ecology and social environment (Wahler, 1975; Willem, 1974). Intensive individualized measures of behavior of large numbers of clients are prohibitively expensive. However, multicategory behavioral observations based on time-sampling procedures (see Arrington, 1943) and administered by nonprofessionals can be conducted at less expense, while maintaining the specificity necessary for rigorous evaluation.

Before methods of direct observation can be recommended for widespread use in applied settings, several methodological and ethical issues must be addressed. Behavior analysts interested in program evaluation cannot expect to replace psychometric tools of known reliability and validity with homemade assessment devices whose psychometric qualities are unexamined (Johnson and Bolstad, 1973; Kahn, 1975). Behavior observation and recording instruments must also be (a) sufficiently inclusive of general classes of behavior to be relevant to the particular program evaluation, (b) simple enough to be used by nonprofessional observers, and (c) sensitive enough to be responsive to differences and changes in programs. This report describes the application of the Behavior Observation Instrument (BOI) as an evaluation instrument for programs in various treatment settings. Information concerning its measurement characteristics is also presented. The material presented is intended to encourage researchers, and especially clinicians, to venture into the difficult but challenging area of program evaluation.

THE BEHAVIOR OBSERVATION INSTRUMENT (BOI)

The BOI was derived from the Behavioral Study Form (BSF) developed by Schaefer and Martin (1966) and from the Location-Activity Inventory (Hunter, Schooler, and Spohn, 1962).

Like the BSF, the BOI consists of two basic types of behavioral categories: *mutually exclusive behaviors* and *concomitant behaviors*. All behaviors that can be observed in a given setting are categorized as one of the mutually exclusive behaviors which may be emitted in conjunction with one or more of the concomitant behaviors. In addition to recording the two types of behaviors, observers also indicate the *location*, or the setting where the behaviors were emitted. Mutually exclusive behaviors are motor behaviors which, in accord with their definitions, cannot occur concurrently—for example, lying down, sitting, standing, walking, and running. A specific episode of behavior would be scored in only one of these categories. Concomitant behaviors, on the other hand, occur in conjunction with mutually exclusive behaviors, and include all other social or individual activities such as conversation, eating, grooming, reading, smoking, watching television, or playing a game. Table 1 details the specific observational classes of the BOI. Operational definitions of BOI categories are provided in DeRisi, Alevizos, Callahan, and Eckman (Note 1).

In its present stage of development, the BOI categories differ from those in previous versions and related instruments. For example, "unusual" or "perseverative" behaviors were added to an earlier version of the BOI because these categories were useful in assessments of environmental or pharmacological interventions in psychiatric settings (see Table 1). The major objective in developing the instrument was to provide a behavioral index of the impact of a treatment program on the activity of individual clients and groups of clients.

The data collected in selected concomitant categories from the BOI have also been com-

Table 1
The Behavior Observation Instrument Coding Categories^a

<i>Mutually Exclusive Behaviors</i>		
1. Walking	4. Sitting	5. Lying down
2. Running	a. Eyelids open	a. Eyelids open
3. Standing	b. Eyelids closed	b. Eyelids closed
<i>Concomitant Behaviors</i>		
6. Drinking	14. In conversation with staff member or visitor	
7. Smoking	<i>v</i> TP is talking to the other person	
8. Eating	<i>nv</i> TP is listening	
a. Meals	15. Inappropriate behavior	
b. Other than meals	a. Perseverative/unusual: motor	
9. Chewing (gum or nails)	b. Perseverative/unusual: verbal	
10. Grooming	16. Help plan, prepare lunch/	
a. Oneself	Help with groceries and shopping/	
b. Another person	Making coffee	
c. Being groomed	17. Cleaning or tidying-up	
11. Reading/Writing/Looking through magazine or newspaper	18. In individual/occupational therapy	
12. Recreational activities	19. In a group meeting	
a. Group activities	a. Actively participating	
b. Solitary Recreation	b. Not active but attentive	
13. In conversation with another patient	c. Apparent inattention	
<i>v</i> Target Patient (TP) is talking	20. Other (e.g., crying or laughing)	
<i>nv</i> TP is listening to another patient	21. Not present	
	a. Away on field trip/assignment	
	b. Unavailable or unknown location	
<i>Locations</i>		
	<i>State Hospital</i>	<i>Day Treatment Center</i>
DR	Day Room	DR Day Room
NS	Nursing Station	OS Staff Offices
DM	Dormitory	OT OT Room
C	Corridors	C Corridors
B/S	Bathroom or Shower	K Kitchen
DI	Dining Room	DI Dining Room
MD	Medication Room	O Other
O	Other	

^aOperational definitions of codes available from the authors.

bined into *summary categories*. For example, the interaction summary category "Social Participation" included any mutually exclusive behavior occurring with concomitant person proximity, eye contact, and conversation (codes 10b or c—grooming and washing another or being groomed and washed; 12a—group recreational activities; 13*v* or *nv*—verbalizing with client, speaking or being spoken to; and 19a or b—attending group meeting/therapy, active or attending). Similar groupings were made for Solitary Behaviors, Inactivity, and Work and

Maintenance Behaviors. The aggregation of codes into summary categories can be accomplished in many different ways to highlight phenomena of special interest to the evaluator or administrator.

Coding Procedures

Observations were recorded on 10- by 15-cm index cards, with the subjects' names, observed codes, and clarifying comments in three columns. At least half of the total population of subjects in each setting were observed. Ten subjects were

observed in The Oxnard Community Mental Health Center's Day Treatment Center (DTC). These subjects were randomly selected from the Center's patient roster of 20 each day that observations were made. Forty subjects were randomly chosen from a female unit roster of 56 at Camarillo State Hospital (CSH). However, the same 40 residents were observed in random order each day in this setting, so that these data might later be compared with other measures. Each subject was observed once for 5 sec during each observation period. Subjects were observed in sequence. If a subject could not be located, he or she was coded "21" (not present) and verification of specific absence was sought after the last subject was observed. Subjects were observed three times each day in the DTC and twice each day in the CSH. The frequency and duration of daily observations were varied experimentally, as described later. The interval between observations of different subjects within a single observation period was permitted to vary in order to allow sufficient time to locate each successive subject. In general, the between-subject interval ranged from 10 to 30 sec.

Selecting and Training Observers

One solution to the problem of obtaining low-cost, reliable data has been the recruitment of students and nonprofessionals as observers. Five types of observers have been used. In decreasing order of cost, they are staff research assistants, college students, high-school students supplied (and paid) by a county summer work program, and minority group high-school students from the federal Neighborhood Youth Corps program. These young people were selected for apparent maturity or judgement and for their willingness for precisely executing the mechanics of the task. These nonprofessionals were each given a training session on regulations and policies concerning confidentiality and patient privacy before being allowed to enter the clinical setting. All of them signed statements certifying they had attended and understood the content of this session. Trainees were given instruction

manuals and time to become familiar with the operational definitions. Next, they were engaged in practice sessions in the classroom and in a variety of community settings (*e.g.*, restaurants, shopping centers). Once in the treatment setting, trainees reached the observer agreement criterion of 85% within the first hour of practice.

Observer Agreement

In the assessment of observer agreement, two independent observers at least 1.5 m apart recorded behavior simultaneously, so that continuous agreement coefficients could be tabulated. One observer signalled the beginning and ending of an interval by lightly tapping his or her clipboard with a pencil.

The calculation of observer agreement in BOI observations used the method of per cent effective agreement (Hawkins and Dotson, 1975). This statistic is the ratio of mutually agreed-upon occurrences. Per cent effective agreement was evaluated in both settings. During initial phases of development, 10 subjects were randomly selected and ordered from the DTC's roster, with four observation periods scheduled each day. The observation periods were 10:00 to 10:30 a.m., 11:00 to 11:30 a.m., 1:00 to 1:30 p.m., and 2:00 to 2:30 p.m. In the CSH, two to 15 observations were taken each day from 8:00 a.m. to 9:00 p.m.

The coded observations of four experienced observers paired with four relatively naive observers were compared, using a stringent act-by-act method of computing the per cent effective agreement. The eight observers were rotated in random pairs across all observations. Only when every behavior recorded in a 5-sec interval agreed between observers was the observation scored as an "agreement". Agreement was calculated for more than 400 individual 5-sec observations. The per cent of effective agreement ranged from 84.6% to 100%, with a mean of 94.9%, in the DTC, and ranged from 84% to 97%, with a mean of 93%, in the CSH.

Per cent effective agreement has been criticized recently because it fails to account sta-

tistically for the probability that two observers can agree by chance alone (Costello, 1973; Hartmann, 1977). The suggested alternative, statistical coefficient *Kappa*, has been used to estimate scorer reliability for nominal scales (Cohen, 1968; Fleiss, 1973) and can be adapted for estimating observer agreement in observational data (Costello, 1973). This statistic is a ratio of mean observed disagreements to the mean of chance expected disagreements. Because certain categories of the BOI are more likely to be used in a given setting, Kappa coefficients were computed to account for chance agreement. The computations were performed on data derived from over two weeks of observation (Derisi *et al.*, Note 1) in the hospital setting. The Kappa coefficients ranged between 0.72 and 0.87 across observation days.

Generalizability of Data Samples

The representativeness of data from BOI observations was determined with reference to Cronbach's generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1973). In this approach, reliability and validity are reconceptualized as the generalizability of a set of measurements to any relevant "universe" of interest; *e.g.*, in behavior assessment, the generalizability or representativeness of a set of observations is, in large part, a function of adequate sampling within the universes of (a) observers or scorers, (b) items or behavior categories, (c) settings, and (d) times of observation (see Cone, 1977).

To determine the representativeness of BOI observations, five observers alternated two at a time in recording and tabulating data. The BOI's 21 categories of behavior and activity provided a broad range of item composition. Clients were observed in a variety of locations within two treatment settings to ensure that the base rates obtained fully reflected these settings. Observers located each client in a pre-ordered, random sequence. This often required movement between different areas in the setting. In the hospital, these included the cafeteria, the hospital grounds, and group therapy.

The representativeness of times of observation was experimentally determined in two ways: by varying first the length of the observation interval and then the frequency of observations.

Observation interval. The goal of making the BOI an efficient assessment procedure involved making the observation interval as brief as possible without sacrificing the representativeness of the data sampled. In the hospital setting, observers recorded each client's behavior for 30 sec twice a day for two weeks. Data were recorded using a cumulative interval scale in which observers coded behavior after 1 sec, then again after a cumulative total of 5, 15, and 30 sec. A 30-sec observation interval was chosen as the longest reasonable time that each of 40 clients could be observed in one session. To conserve space, two mutually exclusive and concomitant categories are presented. Figure 1 illustrates the pattern of results. The results are expressed as the per cent of data yielded in 1, 5, and 15 sec when compared to a 30-sec criterion (100%).

Separate analyses of variance (categories $\times$ intervals) were performed on the mutually exclusive and the concomitant classes of behavior. Only the concomitant categories showed reliable differences across the four observation intervals, $F(3,720) = 18.75, p < 0.001$. A test of simple main effects identified significant differences in the four observation intervals for three concomitant categories: client verbalization with client (VP), $F(3,504) = 19.12, p < 0.001$; unusual and perseverative behaviors (UP), $F(3,504) = 13.97, p < 0.001$; and verbalization with staff (VS), $F(3,504) = 3.60, p < 0.050$. A Newman-Keuls analysis showed that 1-sec observations differed significantly ($p < 0.050$) from both 15- and 30-sec observation intervals across the three categories. In all but two comparisons (VP and UP), 5-sec observations yielded essentially the same data statistically as 15- or 30-sec observations. There were also no statistically significant differences in the proportions of behavior sampled by 5 or 15 sec when 30 sec was used as the criterion ($z = 1.60$ and

Fig. 1. The per cent of behavior recorded during a 30-sec observation interval, which was sampled in the first 1, 5, and 15 sec of observation summed across five days using two concomitant and two mutually exclusive category composites of the Behavior Observation Instrument.

0.77, $p < 0.050$). However, a 1-sec observation interval yielded significantly different proportions when compared to either 5-, 15-, or 30-sec observation intervals ($z = 28.11, 29.63$, and $29.95, p < 0.050$). Thus, in all but two BOI categories, the data at 5-sec intervals corresponded to data from longer intervals; *a 5-sec observation interval is therefore both efficient and representative* when compared to a 30-sec criterion. The categories VP and UP yield reliable data with a 15-sec observation interval. However, when these variable rate activities are sampled across days, the sampling frequency of twice per day with 5-sec intervals was shown to yield representative data.

Frequency of observation. With reference to both efficiency and generalizability, the authors next sought to determine how infrequently individuals would need to be observed to reflect stable base rates in their daily behavior. In the state hospital, long-term residents (chronic psychotics) were observed 12 hr a day during residents' active morning, afternoon, and evening hours. These "marathon" samplings (or criterion samplings) involved continuously rotating observations across each of 40 residents over a five-day period. This resulted in at least 15 observations per resident per day. Initially, each client was observed for arbitrary intervals of 5 and 15 sec during each of the minimum of 15 daily samples.

The 15-per-day marathon observations were then compared with independently sampled observations conducted concurrently by two additional observers over a five-day period. These independent observations (with both 5-sec and 15-sec intervals) were made only twice each day, once in the afternoon and again in the morning or evening (see Figure 2). To evaluate sampling adequacy statistically, individual categories from the marathon observations were ordinaly ranked for 40 clients and analyzed across five days' observation. These marathon data were first compared to the ranks of categories from observations taken at 5- and 15-sec intervals sampled twice a day for the same clients; the rank order

coefficients for each day across were all above 0.89. A second comparison was made by randomly sampling one-third and two-thirds of the marathon observations and correlating these data with the total marathon data. Rank order coefficients ranged between 0.96 and 0.99. Results were not analyzed for each rotating observer, although observer agreement remained above 86% throughout these time studies.

BOI observations were then conducted for four successive weeks to assess stability over days. The per cent of behavior in each of the 21 categories was found to be stable over time (unpublished data, Alevizos and Callahan) and consistent with the proportion of behavior in each category from the two previous studies. These data suggest that within inpatient settings with characteristically low levels of activity, BOI observations conducted twice a day adequately reflect resident activity.

Comparisons among observation methods in the Community Mental Health Center's Day Treatment Program yielded similar findings: using a 5-sec observation interval, 15 clients were observed on a continuous schedule, averaging approximately three observations per client per hour. Clients were observed 18 times during the program's 6-hr day. Observations were made over 15 days. The clients served by this Center represented the broad spectrum of adult psychiatric clients encountered in community mental health settings. About one-third were "reconstituting" psychotics, one-third were characterized as having depressive or anxiety states, and one-third were identified as substance abusers, mild retardates, or sociopaths. They differed markedly from institutionalized state hospital residents by their higher activity levels and by the quality and variety of their social and instrumental behaviors. The levels of psychotropic medication prescribed at this Center were also significantly lower than those used in the state hospital.

The results were analyzed by sampling randomly from the 18 observations per client per day to simulate observations made twice per hour, once per hour (six per day) and every other

Fig. 2. The per cent of total behavior recorded using a criterion of 5-sec observations which rotated continuously between 40 patient-residents across 12-hr days for five days compared to separately sampled 5- and 15-sec observations made twice per day on each of eight concomitant categories of the Behavior Observation Instrument.

hour (three per day). These data were compared to the means of data derived from three observations per hour. The concomitant behavioral categories of the BOI were analyzed in two summary categories—Social Participation and Isolate Behavior. Although day-to-day variability increased somewhat as the frequency of observations decreased from three per hour to one every other hour, mean values across the 15 days did not differ appreciably. The mean percentage of behaviors coded as "Social Participation" varied from 58.4 to 60.6. The mean percentage of behaviors coded as "Isolate Behavior" varied from 33.0 to 34.3. Therefore, as the BOI was used in the Oxnard DTC program, one observation per hour was adopted as a standard for evaluating program change.

Reactivity

A frequent objection to direct observation methods is that the observation process itself may change the behaviors being observed, resulting in anomalous data. This issue has received considerable attention in recent years (Hagen, Craighead, and Paul, 1975; Romanczyk, Kent, Diamant, and O'Leary, 1973) and will undoubtedly continue to be of concern.

Callahan and Alevizos (Note 2), using the BOI, investigated reactivity on a hospital ward for chronic clients and found no evidence of habituation in the residents' data over four weeks' time, nor any significant effects due to the gender of the observers.

Manipulation of observation frequency, from twice per day to continuous observation, showed only a slight increase in observations of clients sitting or lying down with eyelids closed—a possible reactive effect. The frequency of "eyelids closed" did not return to baseline levels when observations were again reduced to twice per day.

When the BOI was modified to assess staff behavior, the results related to reactive effects were inconclusive. When staff at a mental health center's inpatient unit were informed they would be compared to staff at another center, they inter-

acted more frequently with their clients and were more punctual in the operation of their program (Lieberman, DeRisi, King, Eckman, and Wood, 1974). When staff behaviors in an inpatient research unit were compared under overt and covert observation conditions, no significant differences were found in their social, administrative, or work behaviors (Williams, Wallace, Lenhardt, Lafey, Alevizos, and Lieberman, Note 3). The presence of reactive effects may be related to any number of variables, including setting characteristics, information provided to client-subjects, and their expectations and prior experiences with direct observation and program evaluation.

THE BOI IN PROGRAM EVALUATION

STUDY I: COMPARISON OF TWO INPATIENT TREATMENT PROGRAMS

If the BOI is to be used in program evaluation, it is necessary to demonstrate that it is sensitive enough to reflect important differences in behavioral characteristics of the programs. Therefore, two inpatient settings were chosen to test the power of the BOI to discriminate between different programs of the same general type. These inpatient facilities were dissimilar in almost every respect, especially in the activity levels of clients.

Setting

Subjects were observed on a ward for chronic female clients of a large state hospital in Southern California. This facility presents an exceptionally pleasant exterior aspect, resembling a university campus. Its interior spaces differ little from those of other large institutions built during the 1930s. Another group of subjects was observed at the inpatient unit of the community mental health center in Ventura, California, a town located 100 km north of Los Angeles. The facility is unique, in that its design deemphasizes institutional characteristics in its use of multipurpose areas and its open interior space.

Subjects

All 40 persons residing in one ward of the state hospital for "chronic" female clients were observed during this study. Their ages ranged from 28 to 65 yr and most were diagnosed as psychotic. Some had been hospitalized more than 20 yr.

All 19 persons residing in the inpatient service of the community mental health center were included in this study. Approximately half of these subjects were male. Ages of the subjects ranged from 19 to 62 yr. Their psychiatric diagnoses ranged from psychosis to adjustment reaction to adult life. Clients stayed in this facility from three days to two weeks. Alcohol and drug abuse cases were not admitted to this unit.

Procedure

All clients in both settings were observed for 10 days on a time-sampling schedule of three observations per client per hour. All clients in both settings were observed during the 8 hr of the day shift. Two observers were employed for all observations. Observers were under almost continuous scrutiny of research staff during this and the following study. For each observational hour, clients were observed in a pre-ordered, random sequence determined by the use of a table of random numbers. Each client was located and observed for 5 sec followed by a 30-sec interval in which codes were recorded and the next client located.

Observer agreement. Observer agreement was assessed in 25% of all observations taken at both facilities. Per cent effective agreement was computed in this study for all possible pairs of the four observers, and ranged from 92% to 100% across all categories of the BOI.

Data aggregation. All BOI codes were sorted into the following four summary categories: the behavior codes that included some form of social interaction were combined to form the "Social Participation" category; observations of behaviors lacking social content were included

in the "Nonsocial Behavior" category; behavior codes involving activities such as maintenance, cleanup, or meal preparation were classified as "Work"; and behavior codes classified as "Other" included "away on field trip", "unavailable for observation", and "eating lunch". Work and Nonsocial Behavior were mutually exclusive. If a client was observed working alone, the observation was aggregated as Work but not Nonsocial Behavior, since the latter category is intended to represent behaviors indicating a lack of involvement in day treatment activities. When Work and Social Participation were observed, both were counted in the summary categories.

RESULTS

The results of BOI observations in the two facilities are given in Table 2. The chronic ward population was characterized by very low levels of behaviors involving social interaction. Observations that reported these behaviors were found one-tenth as often as on the inpatient unit of the community mental health center. The social participation figure of 3.5% is the lowest recorded for any setting in which the BOI has been used to date. One explanation may be that the higher-functioning clients on the "chronic" ward spent a large portion of the day away from the ward, either because they had grounds privileges or because they had work assignments elsewhere in the hospital complex. Thus, the clients

Table 2

Summary categories from BOI observations in two inpatient settings: per cent of total observations.

	Social Partici- pation	Non- social Be- havior	Work	Other Be- haviors
State Hospital—				
Female Chronic Ward	3.5	67.3	2.1	25.1
Community Mental Health Center—				
Inpatient Ward	34.4	44.7	3.6	17.1

who remained on the ward were a select group. Fully 80.2% of the "Other" behaviors (Table 2) were "off-ward" and "unavailable for observation". Similarly, clients at the community inpatient facility were encouraged at times to leave briefly in order to assist in making arrangements for their postrelease adjustment or to engage in community-oriented activities for therapeutic purposes.

STUDY II: CHANGES IN A DAY TREATMENT PROGRAM

To determine the sensitivity of the BOI to specified changes in the DTC program, behavioral comparisons were made under differing schedules of programmed activities. Observations had been taken before the DTC program had been changed from a traditional milieu therapy model to an educational workshop model emphasizing behavioral procedures. While the full schedule of the new workshop model was in effect, further observations were taken. Programmed activities during this phase accounted for 22 hr of 26 available clinic hours. Clients were involved in program 85% of the time they spent at the clinic.

When it became desirable to revise the workshop program, two-thirds of the workshops were temporarily cancelled. During this 10-day period, staff were asked to return to the traditional milieu approach. Programmed activity was reduced from 22 hr to 8 hr (30.8%) of the 26 hr available weekly.

In reducing the programmed activity so drastically, a 10-day *withdrawal from programming* was accomplished. It was desirable during the withdrawal phase to analyze separately data collected during workshops and during those parts of the day when workshops were cancelled. In no-workshop hours, clients were allowed to determine their own activities, presumably taking advantage of the therapeutic milieu of the treatment setting. The data from the withdrawal phase are presented to depict client behavior during workshops and during the times when workshops were cancelled.

Observer agreement. Six individuals, unaware of the nature of the study, were trained as observers to use the BOI (see Selecting and Training Observers, above) using a per cent effective agreement method on individual ungrouped categories. Observer agreement among six observers did not fall below 90% during this experiment.

RESULTS

The effects of manipulation of degree of program structure on four general response classes of the BOI are given in Table 3. Table 3 shows that "Social Participation" behaviors declined, corresponding to a decrease in programming during the withdrawal phase. "Non-social" behaviors increased during this phase. A further breakdown of the data in the withdrawal phase showed that during the time when the workshops were scheduled, "Social Participation" behaviors accounted for 66% of the total, with "Nonsocial" behaviors accounting for 27%. "Work" and "Other" behaviors were essentially unchanged.

In the withdrawal phase when all workshops were cancelled, "Social" and "Nonsocial" behaviors returned to the levels previously recorded when a "traditional milieu therapy" program was in effect at the center (Lieberman *et al.*, 1974).

When the program was reinstated with fewer workshops than in the full schedule, "Social Participation" and "Nonsocial Behavior" observations stabilized at 58% and 32% respectively.

PRACTICAL CONSIDERATIONS: TIME AND DOLLAR COSTS

In most applied settings, staff typically report too many demands on their time. Formal, direct observation is not a valued activity in such settings, and is seldom included in formal job descriptions. When performed, it is usually done in addition to ongoing clinical responsibilities or as part of a time-limited research or demonstration project with independent funding. Additional cost and diversion of professional staff

Table 3
BOI Summary Categories During Four Experimental Conditions^a:
Per cent of Total BOI Observations

	<i>Structured Program Time^c</i>	<i>Social Participation</i>	<i>Nonsocial Behavior</i>	<i>Work</i>	<i>Other Behaviors</i>
Traditional Milieu Therapy ^b	47.5	29	62	9	0
Educational Workshops ^b (Full Schedule)	84.6	68	17	7	8
Educational Workshops (Withdrawal Phase)	30.8	59	33	3	5
Workshops Scheduled		(66)	(27)	(3)	(4)
Workshops Cancelled		(30)	(57)	(3)	(9)
Educational Workshops (Reduced Schedule)	69.2	58	32	4	6

^aThe data are presented as mean per cent of social participation, nonsocial behavior, work, and other behaviors (eating meals, away from facility on field trip, absent) under six different program conditions at the Oxnard Day Treatment Center.

^bLiberman, King, DeRisi, Eckman, and Wood, 1974.

^cPer cent of available clinic hours per week.

time are most often cited as the major obstacles to the use of direct observation methods. As previously suggested, the use of nonprofessionals has been one method used to offset additional costs. Various incentives can be offered to such observers. Course credit was offered to college students, but the personal attention and supervision of professional staff was found to be sufficient for maintaining good on-task behavior. The time spent on this type of supervision did not exceed that normally needed for the supervision of student interns. Observers usually attended interdisciplinary team meetings and their opinions were sought by the professional clinicians. They rapidly gained status as valued team members with special expertise and information.

Training of observers was streamlined to reduce the amount of professional involvement. There is now little professional involvement in the first phase of training, which lasts 1.5 hr, and training to the criterion performance level takes approximately 4 hr.

The cost of each observation by these workers was calculated to be five cents each time one client was observed. This cost estimate included staff supervision time and materials and assumes

a schedule of one observation of a set of 10 clients per hour.

With a trained observer in a setting with 20 active clients, the entire observation task for one 6-hr clinic day requires approximately 4.5 hr. This includes preparation of schedules and materials, 45 min—the observations themselves consuming 67 min using a 5-sec observation interval and 30-sec between-client intervals; transcribing and aggregation of data on worksheets, 120 min; and storing data cards and materials, 40 min. This estimate assumes a completely manual system; the use of automatic equipment for data entry and storage could reduce the time required for the preparation, data aggregation, and storage functions.

DISCUSSION

The BOI is one method of applying the direct observation methods that were developed for behavioral research to program evaluation. This report demonstrates its applicability and practicality. Because of its relatively simple design and procedures, manpower requirements were found to be within the capability of many programs.

Acceptable observer agreement was rapidly attained across observers of differing backgrounds.

Investigations into the optimal length of observation interval, frequency of observations per hour and per day yielded information useful in maintaining these dimensions of the process, so as to minimize wasted effort. Such developmental investigations should be accomplished whenever a new behavioral observation method is to be applied or when previously developed methods are applied to novel situations.

The BOI was shown to discriminate between inpatient programs (Study I) and to track the effects of changes in a day treatment program (Study II). In the latter study, frequency of social-interactive behavior was found to follow the level of programming provided by staff. From a social learning standpoint, it is reasonable to expect that the amount of appropriate, interpersonal behavior of clients would vary with the amount of staff-client contact.

More detailed and complex analyses of BOI data have been shown to yield useful program evaluation information. Callahan, Alevizos, Teigen, Newman, and Campbell (1975) observed the behavior of 17 persons on a state hospital ward for clients labelled "chronic" in order to determine the effects of a single administration of major tranquilizers (at hour of sleep), compared to the usual multiple-dose schedule. Contrary to the anxious predictions of ward staff, the only reliable change in client behavior during the experimental condition was an increase in the number of clients who had their eyes open; no increases in low-frequency behaviors or other effects were noted in the nursing log. The change in eyes open was far more subtle than those reported in the present report. It may not have been noted if a global assessment tool had been employed, since there were no other behavior changes recorded. One inference that could be made for purposes of program planning is that the single-dose medication schedule left clients more alert and perhaps more responsive to an increased level of program intensity. Single-dose schedules are also less costly in terms

of drugs and nursing personnel time, and free nursing staff for other constructive contacts with clients.

In the development and evaluation of service programs, direct observational methods like the BOI may provide less biased and more detailed descriptions of behavior than retrospective ratings or questionnaires alone (*cf.* Wahler and Leske, 1973). However, as an indicator of the range, distribution, and frequency of social, instrumental role, and symptomatic behaviors, the BOI provides a necessary but not singularly sufficient measure of a program or a therapeutic environment. Optimally, self-report and rating scales should be empirically combined with direct observations to provide a multidimensional assessment of a service delivery system (see Johnson and Bolstad, 1973; Lentz, Paul, and Calhoun, 1971; Paul, McInnis, and Mariotto, 1973). Methodologically rigorous observation systems like the BOI can be useful in evaluating the validity of these other "indirect" measures of client behavior (*e.g.*, Mariotto and Paul, 1974). Observation instruments like the BOI, which focus on the measurement of categories of behaviors, have other specific limitations. The behavior categories of the BOI ignore the content or meaning of social interaction, which may prove vital to individual assessment if not for program evaluation. For example, conversation between patients is always coded the same, yet the content might be argumentative, affectionate, or informative—all would simply be coded as conversation. More generally, the BOI records behavior in a value-free context, so that inferences regarding the desirability or undesirability of the recorded behavior require additional judgements by patients, their families, their therapists, or by reference to social norms. Determining the therapeutic quality of motor activity *versus* inactivity and social participation *versus* isolation involve criteria outside the scope of the measurement instrument. Motor activity and social participation, for example, have quite different value loadings on a ward of patients with acute myocardial infarctions, as con-

trasted with a ward of chronic mental patients. Similarly, the amount of social interaction noted may be "good" or "bad" depending on the kind of patient (depressed *versus* manic) and the content, coherence, and quality of the conversation and interaction.

As a time-sampling instrument, the BOI is relatively *insensitive* to low base-rate behavior. Certain infrequent but high-intensity and annoying behaviors, which may be instrumental in precipitating hospitalization or influencing discharge, are usually missed by time-sampling procedures. Instances of aggression, for example, have rarely been recorded during an observation period. Low-frequency behaviors should be recorded by staff as they occur in the natural setting.

Increased requirements for accountability for clinical services make it necessary to determine whether treatment/training programs have beneficial effects on client behavior, first within and later outside the treatment setting. Direct observation of behavior across all clients can provide this information. The studies described in this report demonstrate that assessment by direct observation, which is generalizable, can be accomplished at low cost, with a limited amount of observer training. In addition, the data provide timely, easily interpreted information for administrators and clinicians. The use of direct observation is not beyond the capability of many clinical programs.

REFERENCE NOTES

1. DeRisi, W. J., Alevizos, P., Callahan, E., and Eckman, T. *Behavior Observation Instrument: instructional manual*. Unpublished manuscript, 1976. (Available from William DeRisi, California State Department of Health, 714 P Street, Sacramento, California 95814.)
2. Callahan, E. J. and Alevizos, P. N. *Reactive effects of direct observation of patient behaviors*. Paper presented at the 81st Annual Convention of the American Psychological Association, Montreal, Canada, August 1973.
3. Williams, S., Wallace, C., Lenhardt, S., Lafey, D., Alevizos, P., and Liberman, R. *Staff reactivity to evaluate observations of unspecified vs. specified*

behaviors. Unpublished manuscript, Camarillo Neuropsychiatric Institute Research Program, Camarillo State Hospital, 1977.

REFERENCES

- Arrington, R. E. Time sampling studies of social behavior: A critical review of technique and results with research suggestions. *Psychological Bulletin*, 1943, **40**, 81-124.
- California State Legislature. *Assembly Bill 2649*. Sections 5651 and 5656. Sacramento: California State Assembly, 1971.
- Callahan, E. J., Alevizos, P., Teigen, J., Neuman, H., and Campbell, M. Behavioral effects of reducing the daily frequency of phenothiazine administration. *Archives of General Psychiatry*, 1975, **32**, 1285-1290.
- Chu, F. D. and Trotter, S. *The mental health complex, Part 1: Community mental health centers*. Washington, D.C.: Center for Study of Responsive Law, 1972. P. 11 (58-59).
- Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, **70**, 213-220.
- Cone, J. D. The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 1977, **8**, 411-426.
- Costello, A. J. The reliability of direct observation. *Bulletin of the British Psychological Society*, 1973, **26**, 105-108.
- Cronbach, L. J., Gleser, G., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measurements*. New York: Wiley, 1973.
- Fleiss, J. L. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, **33**, 613-619.
- Hagen, R. L., Craighead, W., and Paul, G. Staff reactivity to evaluative behavioral evaluations. *Behavior Therapy*, 1975, **6**, 201-281.
- Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 1977, **10**, 103-116.
- Hawkins, R. P. and Dotson, V. A. Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp and G. Semb (Eds), *Behavior analysis: areas of research and application*. Englewood Cliffs: Prentice-Hall, 1975. Pp. 359-376.
- Hunter, M., Schooler, C., and Spohn, H. The measurement of characteristic patterns of ward behavior in chronic schizophrenics. *Journal of Consulting Psychology*, 1962, **26**, 69-73.
- Johnson, S. M. and Bolstad, O. Methodological issues in naturalistic observation: some problems and solutions for field research. In L. A. Hammer-

- lynck, L. C. Handy, and E. J. Mash (Eds), *Behavior change: methodology, concepts and practice*. Champaign, Illinois: Research Press, 1973. Pp. 7-67.
- Kahn, T. C. On behavior assessment. *American Psychologist*, 1975, **30**, 520.
- Lentz, R. J., Paul, G. L., and Calhoun, T. R. Reliability and validity of three measures of functioning with "hard core" chronic mental patients. *Journal of Abnormal Psychology*, 1971, **28**, 69-76.
- Liberman, R. P., DeRisi, W., King, L., Eckman, T., and Wood, D. Behavioral measurement in a community mental health center. In P. Davidson, F. W. Clark, and L. A. Hammerlynck (Eds), *Evaluating behavioral programs in community residential and school settings: proceedings of the Fifth International Banff Conference on Behavior Modification*. Champaign, Illinois: Research Press, 1974. Pp. 103-139.
- Mariotto, M. J. and Paul, G. L. A multi-method validation of the inpatient multi-dimensional psychiatric scale with chronically institutionalized patients. *Journal of Consulting and Clinical Psychology*, 1974, **42**, 497-508.
- Paul, G. L., McInnis, T. L., and Mariotto, M. J. Objective performance outcomes associated with two approaches to training mental health technicians in milieu and social learning programs. *Journal of Abnormal Psychology*, 1973, **82**, 523-532.
- Romanczyk, R. G., Kent, R., Diament, C., and O'Leary, K. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, 1973, **6**, 175-184.
- Schaefer, H. H. and Martin, P. Behavioral therapy for apathy of hospitalized schizophrenics. *Psychological Reports*, 1966, **19**, 1147-1158.
- Wahler, R. G. Some structural aspects of deviant child behavior. *Journal of Applied Behavior Analysis*, 1975, **8**, 27-42.
- Wahler, R. G. and Leske, G. Accurate and inaccurate observer summary reports. *Journal of Nervous and Mental Diseases*, 1973, **156**, 386-394.
- Willems, E. P. Behavioral technology and behavioral ecology. *Journal of Applied Behavior Analysis*, 1974, **7**, 151-165.

Received 23 February 1977.

(Final Acceptance 20 December 1977.)