A BRIEF REPORT ON MENTALLY RETARDED RESIDENTS AS BEHAVIORAL OBSERVERS¹

Mensch's (1966) suggestion that the role of professionals in the field of mental health be changed to a largely consultative one has to some extent become a reality (see Guernsey, 1969). This has been particularly true in the use of paraprofessionals in behavior modification; teachers (O'Leary and O'Leary, 1972), psychiatric and child-care aides (Wagner and Paul, 1970), parents (Zeilberger, Sampen, and Sloane, 1968), and peers (Bailey, Timbers, Phillips, and Wolf, 1971; Surratt, Ulrich, and Hawkins, 1969) have effectively served as behavior observers and change agents.

It is somewhat surprising, given the emphasis on the use of paraprofessionals, that attempts to employ institutionalized mentally retarded residents as either observers or behavior modifiers have been minimal (Craighead and Mercatoris, 1973). The purpose of the present study was to assess the reliability of behavior observations taken by "mentally retarded" residents, and to compare those reliabilities with the reliabilities of similar observations taken by the ubiquitous college "sophomore".

METHOD

Subjects and Setting

Subjects were four female mentally retarded residents of a state school and hospital and two undergraduate students.² The four retarded residents were assigned to work in pairs, the first pair composing Dyad 1 and the second pair, Dyad 2; the college sophomores composed Dyad 3, the comparison dyad. The IQs (WAIS) of

Dyad 1 were 49 and 69, and of Dyad 2, 68 and 69. Their ages ranged from 17 to 40 yr with a mean of 28.5. All four residents could read and write. Each had had some exposure to the checklists from pilot work for the present study, but none had succeeded in using them accurately. None of the subjects had worked in the same pairs before. The residents received tokens (exchangeable for goods in the institution) and quality ratings for participating. The undergraduates, who had had no exposure to the rating form, received partial course credit.

Except for three days for a generalization test for Dyad 1, all ratings were performed in the residential unit of two of the subjects (one from each dyad). Ratings were performed in every room of the unit, except rest rooms and locked offices.

Procedure

Each dyad was trained independently. When a dyad first arrived at the experimenter's office in the residential unit, they were given clipboards and pencils and the nature of the study was explained to them. However, Dyad 3 was led to believe that their ratings were being used to test the effectiveness of massed *versus* spaced trials in checklist learning and, therefore, they were not aware of the other two dyads.

The rating form was a Time-Sample Behavior Checklist (TSBC) (Mariotto and Paul, *in press*; Schaefer and Martin, 1969). The TSBC was a rating matrix (see Table I), allowing for up to 40 independent observations; each person rated constituted an observation. The location and five categories of behavior (see Table II) could be recorded for each person observed. The locations in Category I and the operationally defined classes of behavior in each of Categories II to IV were mutually exclusive, allowing only one entry per cell in each category. Any number of behaviors, within the limits of the categories,

¹This research was supported in part by Hospital Improvement Project Grant No. 51P-70087-3-02 of the Social and Rehabilitation Services, Department of Health, Education, and Welfare awarded to Laurelton State School and Hospital, Laurelton, Pa.

²Appreciation is expressed to Barbara Klein and Faith Witkin, who served as observers.

Table I

Laurelton State School and Hospital Time-Sample Behavior Checklist

Day:	S	M	Tu	W	Th	F	S	_____	Time	_____	Observer	_____
Observation #	I	II	III	IV	V	VI						

Comments:

could be recorded in Categories V and VI for each individual observed.

All dyads were given a TSBC to clip to the front of their clipboards and category sheets

containing only Category I (Table II) to tape to the back. They were then given an introductory trial in which they were shown each room in Category I in the order that rat-

Table II

Laurelton State School and Hospital
Rating Categories for Lynn Cottage Observers

I. <i>Location</i>	V. <i>Social orientation (score all)</i>
1. kitchen hall	1. alone
2. kitchen	2. with residents
3. dining room	3. with people who work here
4. day room	4. with people who do not work here
5. washing room—B10	VI. <i>Activities (Score all)</i>
6. mop room—B1	1. housekeeping (e.g., sweeping, mopping, dusting, scrubbing)
7. porch and benches	2. eating
8. office—101	3. drinking
9. dorm—102	4. talking to others
10. dorm—111	5. watching others
11. 105	6. listening to others
12. 106	7. singing or playing piano
13. *(room with no number)	8. reading
14. back porch	9. writing
15. laundry room—206	10. sewing, crocheting, knitting
16. dorm—204	11. smoking
17. dorm—210	12. playing a game, puzzles
II. <i>Position (Score 1)</i>	13. listening to radio, record player
1. sitting	14. watching T.V.
2. standing	15. combing hair, brushing teeth, washing body
3. lying or sitting on floor	16. talking to self
4. lying on bed	17. screaming or yelling
5. lying on chair	18. swearing, cursing
6. walking	19. mumbling, fast talking (nonunderstandable)
7. running	20. crying
8. dancing	21. rocking (not in a rocking chair)
III. <i>Wake-asleep (Score 1)</i>	22. destroying property
1. eyes open	23. hurting self
2. eyes closed	24. hurting others
IV. <i>Facial expression (Score 1)</i>	25. looking out window
1. smiling, laughing	
2. frowning	
3. not smiling, laughing, or frowning	

ings would always be done. For Categories II through VI, the experimenter modelled each of the behaviors just before the rating of each category.

Training began with only Category I, and for that category, as well as subsequent categories, proceeded in three stages:

1. *"Training"*. During a "training" trial, the subjects and the experimenter simultaneously rated other residents throughout the unit. The first resident encountered was marked Observation Number One on the checklists; if more than one resident were in a particular room, the subjects would agree (by head nod, whispering resident's name, *etc.*) as to which to rate first. They would quietly count to three (approximately 3 sec) while looking at the resident and then independently rate that resident according to the categories on the back of their clipboards. After rating, the experimenter would check the ratings and discuss why they were correct or incorrect, but did not allow the subjects to change their ratings. The subjects then agreed on which resident was to be Observation Number Two and continued rating in the unit following the same procedures. After all ratings, they returned to the experimenter's office and per cent reliabilities (the number of agreement cells divided by the number of agreement plus disagreement cells times 100) were calculated between the subjects and between the experimenter and each subject to serve as criteria for progress through training stages. These criteria reliabilities were calculated over all categories in use on a trial, rather than by single categories. If all three reliabilities were 85% or above, the subjects moved to the "independent" stage; if not, they would perform "training" trials on the same categories until the three reliabilities on a trial met the 85% criterion.

2. *"Independent"*. On an "independent" trial, the experimenter remained in his office while the subjects made ratings. They were given instructions not to interact with each other, except to agree on whom to rate and to count to three. If after the trial their reliability was 85% or

above, they made two more "independent" trials. If on any of the three "independent" trials the reliability fell below 85%, the subjects immediately returned to "training" on those categories. After three successful (85% or above reliabilities) independent trials, the subjects proceeded to the "check" stage.

3. *"Check"*. During a "check" trial, the experimenter again rated with the subjects. During the "check" stage, ratings were not discussed until after returning to the office at the end of the trial. The subjects and experimenter interacted only to agree on whom to rate and to count to three. As in the "training" stage, three reliabilities were computed after a trial. If any of the reliabilities were below 85%, the subjects returned to the "training" stage for those categories. If all three reliabilities were 85% or above, the subjects were given category sheets containing one additional category for their clipboards and proceeded to "training" on the expanded number of categories.

The above paradigm was followed by each dyad until the subjects completed a "check" trial on all six categories with 85% or above reliabilities. Training was then considered completed. Dyads 1 and 2 completed three trials per day, five days a week with 10 min between trials. Dyad 3 massed all their trials into five days but still maintained the 10-min interval between trials. Dyad 3's participation in the study was terminated after training. Dyads 1 and 2 continued rating, after training, to ascertain if their high reliabilities could be maintained. The specific procedures after training were:

Dyad 1. For eight days, Dyad 1 continued rating in the residential unit. The trials were in the order of three "independents" and a "check", three "independents" and a "check", *etc.*, unless a reliability fell below 85%. If so, they returned to the "training" stage for all six categories and followed the paradigm used in training.

These eight days were followed by three days of rating in another residential unit using similar categories, differing only in the locations of

Category I. Subjects followed the same paradigm as in training, except that they began rating with all six categories on their first "training" trial. It was hypothesized that more activities under Category VI would be exhibited in this second unit because of the lower level of residents, and therefore ratings would be more difficult.

Ten days without ratings followed. A one-day follow-up was then performed to see if the subjects had retained their ability to rate without practice. The follow-up consisted of two "independent" trials and one "check" trial and all were conducted in the original unit.

Dyad 2. Only two days of continued ratings followed training. The subjects made three "independent" trials and one "check" trial during this period.

RESULTS

Interrater reliability was calculated by scoring each cell on the rating form as agreement or disagreement. Agreement was defined as exact agreement between observers on the behaviors

rated in a particular cell. Reliability then was computed by dividing the number of agreement cells by the number of agreement plus disagreement cells and multiplying the quotient by 100.

Figure 1 presents the reliabilities, each calculated over all categories in use on that trial, for subjects in Dyad 1. As can be seen, all reliabilities except one (81%) were above 85% during training. The subjects continued to obtain high reliabilities throughout the rating period after training, during the generalization to the other residential unit, and during the follow-up after 10 days of no ratings; in only four instances did ratings drop below 85%, and even then were not below 75% agreement. Overall, the reliabilities ranged from 75% to 100%, with an average reliability of 93.5%.

The reliabilities with the experimenter are not shown but *were comparable* to those presented for the subjects' "training" and "check" trials. The number of residents observed per trial was not significantly correlated with the reliability obtained ($r = -0.13$, $df = 65$).

The reliabilities of Dyad 2 (Figure 2) were above 85% except on an introductory trial

Fig. 1. Interobserver agreement for Dyad 1 under all treatment conditions.

Fig. 2. Interobserver agreement for Dyad 2 under all treatment conditions.

(used to acquaint the subjects with the locations) and two trials³ during training. The overall range of reliabilities was 74% to 100%, with an average reliability of 95%. Again, there were comparable experimenter-subject reliabilities. There was a significant negative correlation between the number of observations per trial and the reliability obtained ($r = -0.30$, $df\ 39$; $p < 0.05$).

The combined reliabilities for both Dyads 1 and 2 ranged from 74% to 100%, with an average reliability of 94%.

³The two low reliability trials followed a negative medical report to one of the subjects. When it was confirmed that she would suffer no dire consequences, her reliabilities returned to their usual level.

Dyad 3's reliabilities (Figure 3), like those of Dyad 1, were above 85% with one exception during training, with a range of 84% to 100% and an average reliability of 96%. There was a significant positive correlation between number of observations per trial and the reliabilities obtained ($r = 0.46$, $df\ 31$; $p < 0.01$).

From Figures 1 and 2 it can be seen that the reliabilities for "independent" trials (mean = 95.8%) for Dyads 1 and 2 were generally higher than those of the "training" (mean = 91.4%) and "check" (91.3%) trials. The reliabilities of Dyad 3 (Figure 3) were more consistent across "training" (95.2%), "independent" (95.9%), and "check" (97.2%) trials.

The reliabilities were also calculated within each category for all check trials for all three

Fig. 3. Interobserver agreement for Dyad 3 under all treatment conditions.

dyads. The ranges and mean reliabilities for each category, by Dyad, were: Dyad 1, I—88 to 100% (mean = 99%), II—67 to 100% (mean = 89%), III—83 to 100% (mean = 95%), IV—79 to 100% (mean = 90%), V—50⁴ to 100% (mean = 87%), and VI—58 to

100% (mean = 76.4%); Dyad 2, I—88 to 100% (mean = 97%), II—80 to 94% (mean = 87%), III—81 to 100% (mean = 96%), IV—80 to 100% (mean = 90%), V—100%, VI—77 to 78% (mean = 77.5%); Dyad 3, I—100%, II—87 to 100% (mean = 97%), III—100%, IV—82 to 100% (mean = 92%), V—93 to 100% (mean = 97%), VI—0.63.

⁴This reliability was based on a uniquely low number of observations, two; the next lowest reliability in this category was 71%.

DISCUSSION

The most significant finding of this study was that mentally retarded residents can be taught to rate, reliably, other residents' behaviors on a TSBC. The method of training worked effectively, though the subjects reported boredom (a frequent complaint of any observers) with the earlier and easier categories. There may be equally effective alternative training strategies; in our work, we have shortened the procedure to one "training", one "independent", and one "check" trial for each category.

Within the limitations of this study, it appears that once retarded residents are taught to rate in one unit or cottage, they may rate in other similar units with minimal instruction and little or no loss of reliability. Additionally, such residents may be expected to be able to miss days of ratings and maintain highly reliable behavioral ratings. Although the follow-up in this study was only 10 days, subjects have, in a similar project now being conducted, missed up to 30 days with virtually no decrement in inter-rater reliability.

The present data generate several questions. The first (and most significant) is whether or not mentally retarded residents can function *independently* as behavioral observers. The frequently noted advantage of saving staff time by using mentally retarded residents as paraprofessionals has received little support (Craighead and Mercatoris, 1973). In fact, in most studies the experimental staff probably could have done the entire task as effectively and perhaps as efficiently as they did by training mentally retarded paraprofessionals. The present study did demand a great deal of time in training and check runs, and while it appears from the data that occasional "check" runs are essential, it is believed that one supervisor could train, occasionally check, and gradually build up an effective corps of mentally retarded residents, who could provide ongoing institution-wide TSBC data. The observers should be carefully screened on the basis of responsibility

(e.g., gets to tasks on time, does not cheat on tasks) and ability to work under stress without frequent "emotional" behaviors.

Another finding was that the retarded subjects scored higher on the "independent" than on the "check" trials, probably because the subjects were not operating totally independently of each other on "independent" trials. In fact, on at least one trial, a staff member observed the residents to be overtly cheating, one resident copying from another. This may have slightly inflated the reliabilities of some of the independent trials, but the adequate reliabilities on the check trials, when cheating could not occur, indicate that the subjects could reliably and validly (using experimenter as criterion) use the TSBC. Such effects did not appear in the college sophomore ratings. As has been found in other settings, it is likely that the observer's awareness that reliabilities were being computed for their data resulted in improved performance (Reid, 1970; Romanczyk, Kent, Diamant, and O'Leary, 1973). However, it seems unlikely that this awareness of reliability calculation resulted in a differential effect for the mentally retarded as *versus* "normal" observers.

With one retarded dyad, the number of residents observed was inversely related to the obtained reliabilities; just the opposite relationship was found for the college students and no significant relationship was found with the other retarded dyad. It is not possible, therefore, to conclude that there is some relationship between the number of observations per trial and level of rater performance. However, based on the within-category reliabilities, it appears that complexity of the rating category results in decreased reliability. This appears to be a problem for both the retarded and the college-student observers. This argues for the use of reliabilities calculated by individual categories, rather than across all categories, as criteria for progress through training stages; we have utilized this method of training and reliabilities have improved both within and across categories. One should use successive approximations, tailored

to the ability level of his subjects, in training behavioral observers.

There are additional considerations, particularly for the retarded subjects, that warrant further research in this area. One of the frequently noted benefits of such paraprofessional programs has been the "double-change" phenomenon (Guerney, 1969), in which the paraprofessionals' behavior changes as much or more in the desired direction as does the behavior of the resident they are attempting to modify. Such changes may be partially due to modelling because of increased exposure to the regular staff. They may also be due partially to the effects of recording others' behavior and becoming more keenly aware of what behaviors are expected, appropriate and inappropriate. Such ratings and self-monitoring have proven to be effective with other populations, and provide an interesting potential area for research with the mentally retarded (Mahoney and Craighead, *unpublished*).

One final advantage is that such programs may provide a prototypic job experience for high-level residents who may at some point be expected to leave the institution. They are assigned an important, meaningful job that has some purpose. They are expected to work on a fairly rigid schedule and are paid for the completed job. The contingencies may be manipulated by the staff to meet the paraprofessionals' needs in terms of job training.

In summary, the mentally retarded residents did serve effectively as behavioral observers, and, in spite of the problems encountered, the findings hold promise for their increased use as observers and paraprofessionals in behavior modification.

W. Edward Craighead
Michael Mercatoris
Dept. of Psychology,
417 Bruce U. Moore Bldg.
Pennsylvania State University,
University Park, Pa. 16802
Barbara Bellack
Laurelton State School and Hospital
Laurelton, Pa.

REFERENCES

- Bailey, J. S., Timbers, G. D., Phillips, E. L., and Wolf, M. M. Modification of articulation errors of pre-delinquents by their peers. *Journal of Applied Behavior Analysis*, 1971, 4, 265-281.
- Craighead, W. E. and Mercatoris, M. Mentally retarded residents as paraprofessionals: a review. *American Journal of Mental Deficiency*, 1973, 78, 339-347.
- Guerney, B. G. (Ed.). *Psychotherapeutic agents: new roles for non-professionals, parents, and teachers*. New York: Holt, Rinehart, and Winston, 1969.
- Mahoney, M. J. and Craighead, B. K. *Self-control techniques in behavior modification with the mentally retarded*. Unpublished paper read at meetings of American Association on Mental Deficiency, Atlanta, June, 1973.
- Mariotto, M. J. and Paul, G. L. A multimethod validation of the Inpatient Multidimensional Psychiatric Scale with chronically institutionalized patients. *Journal of Consulting and Clinical Psychology*, in press.
- Mensh, I. N. *Clinical psychology: science and profession*. New York: Macmillan, 1966.
- O'Leary, K. D. and O'Leary, S. G. (Eds.). *Classroom management: the successful use of behavior modification*. New York: Pergamon Press, 1972.
- Reid, J. B. Reliability assessment of observation data: a possible methodological problem. *Child Development*, 1970, 41, 1143-1150.
- Romanczyk, R. G., Kent, R. N., Diamant, C., and O'Leary, K. D. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, 1973, 6, 175-184.
- Schaefer, H. H. and Martin, P. L. *Behavior therapy*. New York: McGraw-Hill, 1969.
- Surratt, P. R., Ulrich, R. E., and Hawkins, R. P. An elementary student as a behavioral engineer. *Journal of Applied Behavior Analysis*, 1969, 2, 85-92.
- Wagner, B. R. and Paul, G. L. Reduction of incontinence in chronic mental patients: a pilot project. *Journal of Behavior Therapy and Experimental Psychiatry*, 1970, 1, 29-38.
- Zeilberger, J., Sampen, S. E., and Sloane, H. N., Jr. Modification of a child's problem behaviors in the home with the mother as therapist. *Journal of Applied Behavior Analysis*, 1968, 1, 47-53.

Received 4 May 1973.

(Revision requested 18 July 1973.)

(Final acceptance 14 January 1974.)