12-1982

The Effects of Behaviorally Anchored Rating Scales as Antecedent and Consequent Stimuli upon Instructional Behaviors in a Special Education Setting

Dennis L. Van Hartesvelt
Western Michigan University

Follow this and additional works at: http://scholarworks.wmich.edu/masters_theses
Part of the Industrial and Organizational Psychology Commons

Recommended Citation
Van Hartesvelt, Dennis L., "The Effects of Behaviorally Anchored Rating Scales as Antecedent and Consequent Stimuli upon Instructional Behaviors in a Special Education Setting" (1982). Master's Theses. 1748.
http://scholarworks.wmich.edu/masters_theses/1748

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
THE EFFECTS OF BEHAVIORALLY ANCHORED RATING SCALES AS ANTECEDENT AND CONSEQUENT STIMULI UPON INSTRUCTIONAL BEHAVIORS IN A SPECIAL EDUCATION SETTING

by

Dennis L. Van Hartesvelt

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Master of Arts
Department of Psychology

Western Michigan University
Kalamazoo, Michigan
December 1982
THE EFFECTS OF BEHAVIORALLY ANCHORED RATING SCALES AS ANTECEDENT AND CONSEQUENT STIMULI UPON INSTRUCTIONAL BEHAVIORS IN A SPECIAL EDUCATION SETTING

Dennis L. Van Hartesvelt, M. A.
Western Michigan University, 1982

The present study was an attempt to determine the effects of the Behaviorally Anchored Rating Scale (BARS) upon three distinct instructional behaviors when presented before instructional sessions and when presented after instructional sessions. A multiple baseline design across behaviors was used with four subjects with a reversal element used with one of those subjects. A slight positive effect of BARS presentation upon instructional behaviors occurred. No significant differences were observed between preinstructional and postinstructional BARS presentation. The relevance of the present study is explored in relation to previous BARS studies, along with the implications of the present study.
ACKNOWLEDGEMENTS

There are several persons whose aid and cooperation were essential to the completion of this thesis. I would first like to thank the administration at the Croyden Avenue School of the Kalamazoo Valley Intermediate School District for providing the research facilities for this project. I would also like to thank Joan Lundahl, Michele Polver-Rueber, and Karin Uhlman for providing staff and students as well as Wayne Fuqua, Ph. D., and Norman Peterson, Ph. D., for providing me with academic guidance which set the stage for this thesis. And finally, my most sincere thanks are offered to Dale Brethower, Ph. D., who led me through the confusing and difficult task of turning a few ideas and academic preparedness into a major literary work. Without this help, this task would not have been accomplished.

Dennis L. Van Hartesvelt
INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of “sectioning” the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
VAN HARTESVELT, DENNIS LEE

THE EFFECTS OF BEHAVIORALLY ANCHORED RATING SCALES AS ANTECEDENT AND CONSEQUENT STIMULI UPON INSTRUCTIONAL BEHAVIORS IN A SPECIAL EDUCATION SETTING

WESTERN MICHIGAN UNIVERSITY

M.A. 1982

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ............................................. 11
LIST OF FIGURES .................................................. iv
INTRODUCTION ....................................................... 1
METHOD ............................................................... 16
    Subjects ......................................................... 16
    Setting ......................................................... 16
    Scale Development ............................................ 18
    Experimental Design ......................................... 19
    Data Collection ................................................ 19
    BARS Presentation .............................................. 20
    Reliability ..................................................... 21
RESULTS .............................................................. 22
DISCUSSION .......................................................... 33
APPENDIX A. SD Presentation ...................................... 43
APPENDIX B. Prompting ............................................. 44
APPENDIX C. Consequation ........................................ 45
APPENDIX D. Session Monitoring Form ........................... 46
APPENDIX E. Interactions between Observer and Subject ......... 47
BIBLIOGRAPHY ........................................................ 49
LIST OF FIGURES

1. Instructor Consequation, Prompting, and $S^D$ Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject A . . . . . . . . . . . . . 23

2. Instructor Consequation, Prompting, and $S^D$ Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject B . . . . . . . . . . . . . 24

3. Instructor Consequation, Prompting, and $S^D$ Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject C . . . . . . . . . . . . . 25

4. Instructor Consequation, Prompting, and $S^D$ Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject D . . . . . . . . . . . . . 26
INTRODUCTION

The purpose of this study was to determine on a basic and empirical level whether the Behaviorally Anchored Rating Scale (BARS), which is normally used as a performance evaluation tool, could also serve as a training tool using presentation of the scales as feedback for appropriate instructional behaviors.

Performance evaluation of employees within organizations in the western world has become a prominent topic in contemporary management and psychology journals (Allan & Rosenberg, 1978; Borman, 1978; Keeley, 1978) and in virtually every textbook dealing with industrial psychology (Kane & Lawler, 1979; McCormick & Tiffin, 1974; Wexley & Yuki, 1977). An outline of common subjective performance evaluation tools and techniques has been provided by Wexley and Yuki (1977).

An essay evaluation is the most open-ended of the techniques. Here, the supervisor simply writes, in essay form, a summary of the performance of each subordinate. This form may be more personal than others, but it is time consuming and lacks quantification for employee comparison. In a group appraisal, the supervisor of the employees to be evaluated as well as three or four other supervisors familiar with them meet to discuss their performance. This method tends to eliminate much error in the essay evaluation; but again, employees cannot be compared quantitatively; and it is very, very costly to the organization. In a field review, a trained personnel manager goes into various sections or divisions of the organization to discuss the performance of individual employees with supervisors and sometimes
with the employees themselves. With the information gained, the personnel specialist then writes an evaluation of the individuals' performance which is amended by their supervisor(s). This method cuts down on paperwork by supervisors and is probably more consistent, if not more accurate, than essay evaluations and group appraisals. But once again, employees cannot be quantitatively compared; and it is probably not as accurate as observation-based evaluation procedures.

Employee comparisons involve weighing up one employee against another to determine performance rankings. Alternation ranking is the simplest of the three comparison techniques, but it may lack accuracy. Here, the supervisor determines the highest and lowest performing employees in a particular group and places their names at the highest and lowest points on the scale respectively. Then, he takes the next highest and lowest performers and does the same. This is repeated until the list is complete. In the paired comparison technique, each employee is compared to every other. This can be done on a matrix where the name of the better performer is inserted in the intersection for the names of each pair of employees. Rank can then be determined by simply counting the number of times each employee was preferred. While not performance specific, this is a fairly accurate means of ranking employees and can be used if only rankings are necessary. A forced distribution can be used for large groups of employees where rough discriminations between them will suffice and when it can be assumed that job performance conforms to a normal curve. Five spaces are provided in which the supervisor places the names of employees who perform: (a) in the lowest 10%; (b) in the next 20%; (c) in the middle
40%; (d) in the next highest 20%; and (e) in the highest 10%. The problem here is that job performance often does not conform to normality, especially where there are effective selection and/or training techniques (this would skew the distribution to the left). And while employees fit into neat little categories, no information is gained about performance levels so salary and promotion decisions are difficult to make with this comparative procedure.

The critical incident technique first described by Flanagan (1949) is an observation-based technique. The supervisor keeps a running log for each employee in which he records specific instances of critical contributory and obstructive work incidents. This log then is referred to directly during the performance evaluation. While this is an observation-based technique, it is not quantifiable and does not take into account the normal ongoing behaviors by employees vitally important in attaining effective production levels.

The weighted checklist is one where statements about poor to excellent performance are listed in random order. The score values are undisclosed to the rater who checks every statement on each list that represents the employee's performance (various dimensions of performance areas and attitudes are often included), and the lists are scored by a judge. The statements themselves, however, often give away their values since the raters recognize various performance levels and those raters may avoid checking negative statements. In a forced choice checklist, blocks of four statements about poor to excellent performance are listed where the rater checks one statement which is most like the employee and one statement which is least like the
employee. The statements are then scored and a performance rating can be made. While possibly reducing rater error, little information is gained about specific performance to apply to personnel development.

Three rating scales were outlined, two of which were very similar. In the graphic rating scale, the subheadings "poor," "average," and "excellent" are placed at their respective places from left to right on the scales, each of which has some performance or attitude heading. The rater can then mark any point on the scale (i.e., between average and excellent) to represent the employee's performance. The multistep scale is very similar, except that the rater can mark only in the three to five evenly spaced boxes provided to represent from poor to excellent performance or attitude. The employee's dimension ratings can be compared with other employees' ratings or can be averaged and compared. Wexley and Yuki contend that both of these scales are quite susceptible to rater errors.

The Behaviorally Anchored Rating Scale has probably been the most widely discussed performance evaluation technique in the last two decades. In this Scale, a dimension of performance is rated on a scale which represents poor to excellent performance by specific examples of behavior. This is sometimes called a Behavior Expectation Scale because the anchors (behavioral examples) are often preceded by "The employee can be expected to . . . ." The dimensions are usually represented by three to ten anchors spread out over the continuum of the scale. Each anchor is assigned a particular numerical value that is meaningful only in reference to the entire scale which may be scaled by tenths or by wholes and whose upper limit may be from 2.5 to 10.
At the time of the performance evaluation, the rater checks the anchor in each of about five to ten dimensions which most represents the employee's typical work performance. The performance evaluation can be made numerically so that comparisons between employees can be done along dimensions and along overall averages. The corresponding anchors can also add valuable performance information for personnel development. As a result of this flexibility, the Behaviorally Anchored Rating Scale may be a better evaluation tool than the others.

However, with all these evaluation tools at hand, Williams (1972) claims that there is substantial evidence around the world to suggest that many performance appraisals are a waste of time. He provides evidence that many appraisals are inaccurate; and in many cases, whether the appraisals are accurate or not, the information is not used as a basis for salary raises, promotion, or personnel development. Williams, however, has not given up on personnel evaluation; and he promotes further development of effective performance evaluation.

The Behaviorally Anchored Rating Scale, which is the focus of this study, has been touted as a promising new tool in staff and personnel evaluation, at least on a statistical level (Kearney, 1976). Bernardin and Smith (1981) maintain that the Behaviorally Anchored Rating Scale (BARS) is different from other methods of evaluation qualitatively in that it is designed to standardize the rating and the observation of employees. They point out that with standardized observation, there is a common ground for reliable rating of performance.

The specific methods for BARS development vary somewhat, but the
The author has encountered from five to eight specific steps of BARS development in a nearly exhaustive skimming of the BARS literature. Blood (1974) has outlined three general steps which cover the typical BARS development. First, behavioral incidents of various degrees of job performance dimensions are generated by a pool of people knowledgeable in the job for which the BARS is being developed. Secondly, the incidents are retranslated. This involves reassorting the incidents from a randomized pool into their proper dimensions. In order for incidents or anchors to be retained, they must be reliably placed in their dimensions by a sound majority of the judges. Finally, the anchors are assigned scale values according to the level of performance they represent. The mean of all values assigned is the final scale value of the anchor, and incidents with high variances are discarded. Occasionally, the job dimensions themselves are developed through a generation and retranslation procedure where the dimensions are reassigned from a random pool to groups of anchors. There are numerous cookbook approaches in the literature which deal with BARS development and which may be consulted for further details (Bernardin, 1977; Bernardin, Alvarez, & Cranny, 1976; Bernardin & Smith, 1981; Borman & Dunnette, 1975; Borman & Vallon, 1974; Campbell, Dunnette, Arvey, & Hellervik, 1973; Das, Frost, & Barnowe, 1979; DeCotiis, 1977; DeCotiis, 1978; Dickinson & Zellinger, 1980; Ivancevich, 1980b; Landy & Guion, 1970; Kearney, 1976; Keavney & McGann, 1975; Millard, Luthans, & Ottemann, 1976; Shapira & Shirom, 1980; Smith & Kendall, 1963; Staples & Locander, 1972; Zedick & Baker, 1972; Zedeck, Kafry, & Jacobs, 1978).
The typical performance evaluation without a BARS usually involves a few simple steps. For hourly employees, their supervisors decide how well they perform their specific tasks and what their work attitude is like, and check their attendance and punctuality. For salaried employees, a few objectives are set at the beginning of the year such as in an MBO program; and at the annual evaluation, some evaluation tool may be used to determine overall job performance, and objectives are examined to see if they have been met. The BARS appears to be a superior procedure to these on four points. First, the supervisors who use it are, in most cases, deeply involved in its development which may provide some face or content validity as well as concurrent validity. Secondly, those supervisors are thoroughly familiar with the BARS' conceptions and parameters so they are experts in its use which may provide some reliability in observation and evaluation. Third, the scale dimensions are broken down into behavioral descriptions of most or all performance levels (on a continuum of poor to excellent) as opposed to a one to five rating of how the rater believes the ratee is doing in a great number of attitudes and performance requirements. This provides objectives for the employee and definitions of performance which should help them to avoid undesirable behaviors, therefore, leaving room for desired performance. Finally, the specificity of the tool may make evaluation easier thus promoting more frequent evaluation. Wexley and Yuki (1977) point out that frequent evaluations of performance may help organizations meet their short- and long-term objectives.

There is evidence in psychometric terms to support the superiority
of BARS over other rating scales, such as the Graphic Rating Scale, the Mixed Standard Scale, and the Multistep Scale. The BARS has been shown to increase interrater reliability and to reduce rater bias (Landy & Guion, 1970; Millard et al., 1976; Zedeck & Baker, 1972) and has decreased leniency and rating variability (Das et al., 1979; Shapira & Shirom, 1980; Vance, Kuhnert, & Farr, 1978). Ivancevich (1980b) reports that with the implementation of BARS for professional employees, there was an improvement in attitude about equity in job requirements and clarity in feedback.

BARS has been used in a variety of jobs including nurses (Smith & Kendall, 1963; Zedeck & Baker, 1972), engineers (Landy & Guion, 1970), retail personnel (Campbell et al., 1973; Staples & Locander, 1972), tank crews in the Israeli army (Shapira & Shirom, 1980), college professors (Das et al., 1979; Dickinson & Zellinger, 1980; Vance et al., 1978), bank managers (Schwind, 1978), and sales personnel (Ivancevich, 1980a).

There are, however, increasing numbers of studies which state that the BARS is no better but no worse than other formats, such as graphic rating scales (Keaveny & McGann, 1975), behavior observation Scales (Latham, Fay, & Saari, 1979) and mixed standard scales (Dickinson & Zellinger, 1980). Keeley (1978) claims that the BARS is limited in application to routine tasks since it is anchored to specific instances of behavior and it may not be as flexible as a graphic rating scale.

Atkin and Conlon (1978) outline other possible problems with the BARS. First, there may be overlap or nonexclusiveness of items within each dimension. Second, critical incidents of behavior (extremely
good or poor performance) may stand out and be reflected in the rating process while routine performance is overlooked. And finally, the predisposition of the rater toward ratees may affect his ratings of those people. But these problems would be common to all rating procedures and neither unique to the BARS, nor solved by the BARS.

Kingstrom and Bass (1981) cited several studies which contradict the conclusions of heretofore popular pro-BARS studies. In their review, Kingstrom and Bass show approximately equal numbers of studies finding superiority with the BARS and with the alternate format scales (graphic, summated, mixed standard) in leniency, interrater reliability, ratee discriminability, dimension independence (halo), and validity. A few studies showed no differences in these. The authors' conclusions were that there were negligible differences between BARS and alternate format scales. This was qualified, however, with the fact that most of the alternate formats used for comparison with BARS in these studies were developed from information derived from BARS developmental procedures. This may indicate that the scale development procedure and not the scale itself is the critical variable.

Jacobs, Kafry, and Zedeck (1980) concur with the findings of Kingstrom and Bass (1981). In a review of at least 39 studies dealing directly with BARS comparisons, it was their conclusion that it was neither superior nor inferior to any other format it has been compared to. It seems that the problems with BARS as a rating tool are the same as with any other rating tool and that each may be just as good or bad as the other.

In shedding new light on the advantages of and problems with BARS,
the author points out that all these studies lacked any raw analysis of the effects of BARS application upon employee performance. Wexley and Yuki (1977) contend that "current research suggests that the extra time and effort required to build behavior-based rating scales may not be warranted if used solely for performance ratings" (p. 207).

Blood (1974) has outlined four spinoffs for BARS development. First, as alluded to earlier, the organization can benefit from a scale which can be developed by representatives of various organizational levels, therefore, gaining important perspectives of the job for which the BARS is being developed. Secondly, the information about the job obtained in BARS development can be used to create a valid training program. The trainees would learn what is expected of them and how they would be evaluated. Thirdly, when there is disagreement in the BARS development over the value of anchors or in what dimension they belong, there is a signal that organizational policy does not exist for these performance requirements or it is not being conveyed. If there are problems with policy, they can then be ironed out according to the information gained in the BARS development process. Where policy exists but disagreements occur, the fourth spinoff comes into play. The organization can determine where the breakdown is in communication and deal with it.

In an endeavor to test the parameters of BARS application, the present study was an attempt on a simple, basic, and empirical level to determine whether the BARS method, used in a very limited context, would be functionally related to performance improvements in terms of actual behavior change as a result of BARS presentations alone. The
BARS information provided by ratings accompanied by specific examples of performance can be likened to feedback. Feedback is defined as information about performance, compared to a standard, and used to guide performance. The information about performance here would be the performance rating. The standard would be the highest anchor, and an attainment of this would demonstrate a guidance of performance.

Prue and Fairbank (1981) outline parameters of feedback important in attaining optimum effectiveness in the particular situation where it is to be used. They point out that performance feedback is popular in attempting to create performance improvements and that an extensive data base has proven that providing information to groups and to individuals about the quantity and quality of their performance does improve it. Feedback has certain advantages which enhance its popularity. It is inexpensive to deliver; it is easy to implement since elaborate training is not a prerequisite to accurate administration (in the case of BARS, limited application may make its extensive development prohibitive); it cuts back the use of nonpolicy aversive control procedures; and it is useful where monetary rewards cannot be delivered, as in a union environment.

Four feedback mechanisms have been outlined: mechanical, self-recorded, verbal, and written. Mechanical feedback is that provided by an apparatus, such as a cumulative talley odometer or a videotape of widget assembly. This is an inexpensive form of feedback since it has moderate initial and little ongoing cost and it will never administer an emotional reprimand. Self-recorded feedback is achieved by the employee recording his own performance on an odometer or by
written tally. This is advantageous when the employee has little supervision or when there is no permanent product. Verbal feedback is that provided by a supervisor to an employee on a direct interpersonal level. This is an easy method of feedback delivery; but it may lack consistency, sincerity may not be evident, and if an adverse relationship exists between a supervisor and employee, it may not be at all effective. Written feedback can easily be made consistent. It is a permanent product, and it can easily fit in with any information processing system the organization presently utilizes. The information gained from BARS as a feedback tool also has these qualities.

The content of the feedback can include individual performance as a percentage of the group's performance, comparison of a group's performance with a standard, comparison of a group's performance with previous performance, comparison of an individual's performance with a standard, and comparison of an individual's performance with previous performance. The information contained in an individual BARS presentation can be compared with a standard (the highest anchor) and with previous performance (previous range).

Prue and Fairbank also discussed the effects of feedback upon performance. They point out that feedback effects have in the past almost blindly been characterized as reinforcing. Since response rates were increased, this could be true; but they also point out that other considerations must be made. One of these would be the stimulus control which feedback exerts over behavior, such as when an organization sets new requirements for reinforcement. Performance could be
rule governed where feedback is used for the first time in the organization so there would be no history of contingent reinforcement. Feedback could also alter the organizational environment such that informal discussions would center around performance improvement and nonprogrammed social reinforcers would be available. It could also set up an atmosphere of competition. Finally, they point out that with consequence based on objective data, employees are more satisfied with their work and surroundings which could enhance the feedback system.

The effectiveness of BARS in improving performance was evaluated in the present study by examining its effects upon staff instructional behaviors in a behavior modification oriented special education facility. In the same facility, Shook, Johnson, and Uhlman (1978), in two separate experiments, studied graphing by staff using (a) response effort reduction, instructions, and group feedback (a notice board with a percentage of up-to-date graphs listed), followed by (b) instructions, group feedback, individual feedback (notice of percent of daily graphs which were current), and reinforcement (social praise). The latter procedure appeared to be most effective and most costly, while individual feedback alone appeared to be most cost effective in terms of response effort by experimenters to deliver descriptive praise to each subject compared with one easily assembled chart.

In an attempt to analyze the feedback effects of the BARS, the scale in the present study was presented as both an antecedent and as a consequence. Krumhus and Malott (1980) studied the effects of modelling, feedback before a session, and feedback after a session upon
Descriptive social reinforcement of students by tutors. No difference was observed between immediate (just after session) and delayed (prior to next session) feedback, but there were marked increases from baseline to intervention in both pre- and postsession feedback and some increase from postsession modelling to pre- and postsession feedback in tutor responses. Feedback, both before and after sessions, consisted of a review of a videotape of the subject's most recent session by the observer and the subject, with the subject citing three instances and noninstances of appropriate descriptive social reinforcement.

In a study by Sepler and Meyers (1978), "... verbal instruction alone was shown to be insufficient for teaching the full range of behavior modification skills required to function effectively in applied settings" (p. 198). In this study, paraprofessionals were given a written pretest, instructed in behavior modification skills, and then given a written posttest. They were also observed for appropriate responses throughout the study.

Training was found to be essential in increasing the proportion of specific training behaviors by paraprofessionals during therapy periods (Fabry & Reid, 1978) and in increasing appropriate S^D's while reducing the elapsed time of each task (Carnine & Fink, 1978).

As suggested by Blood (1974), the BARS could serve first as an evaluation tool to determine what skills a therapist demonstrates and lacks and secondly as a training tool by showing the therapist in what way he or she is ineffective and what must be done to attain effectiveness. A trainer could also use this information to differentially reinforce behaviors or appropriate approximations thereof.
which contribute to effectiveness. As pointed out above, such training was found to be important by Fabry and Reid (1978) and Carnine and Fink (1978) in developing behavior modification skills. By surpassing instruction alone, which Sepler and Myers (1978) found insufficient in training behavior modification skills, the BARS could well be a very efficient tool for personnel development.
METHOD

Subjects

The subjects were all employees of a special education facility and were all female. Subject A was a 24-year-old certified teacher who, in this study, worked with three students from her classroom on Distar Math. She had previous experience in Distar Math. Subject B was a 21-year-old instructional aide working with the same three students as Subject A in Distar Math. Subject B had no previous experience in Distar Math. Subjects A and B alternated progressive lessons so that neither ran the same lesson, although lessons often repeated drills from previous lessons. Subject C was a 28-year-old certified teacher who was working with the same three students as above whom she did not otherwise normally work with. In addition, she worked with one of her own students who joined the group of the other three in Distar Language. Subject C had minimal previous experience in Distar Language. Subject D was a 21-year-old part-time paraprofessional working with a different student than those above on Distar Math. She had no previous experience in Distar Math. All of the math lessons for Subjects A, B, and D were conducted from Distar Math Book I. A. All of the language lessons for Subject C were conducted from Distar Language Book I. A.

Setting

Subjects A, B, and C all worked in the same area at different
times. The area included a long table, on either side of which sat the students and the subjects (instructors); a blackboard behind the instructor; and Distar Language or Math materials on the table. Subject D worked in an individual session both, which included a student's desk, a chair facing the student in which the subject sat with Distar Math materials on the desk. The observer always sat in a position somewhere behind the student or students so as not to interfere visually with the subjects.

The students all attended classrooms where behavior modification was the main mode of student motivation. This behavior orientation required consistent behaviors by the instructors in pretask or antecedent stimulus presentation, correction, or prompting and posttask or consequent stimulus presentation. Paraprofessionals and instructional aides such as Subjects B and D normally received in-service training in classroom behavior management including behavior modification in instructional sessions. They were usually monitored bi-monthly in instructional behavior modification. Certified teachers such as Subjects A and C regularly participated in classroom behavior management conferences; but in the facility where they were employed during this study, Subjects A and C received no formal training nor were they monitored in instructional behavior modification. In the past, however, they had both received intensive training in behavior modification in instructional sessions concurrent with previous positions which required this training.
Scale Development

The three BARS scales were developed by the author (see Appendices A-C) from job models previously written by the present author and by other experts within the facility in job model development as well as from the Session Monitoring Form (see Appendix D) which was developed by employees of the facility who were experts in programmed instruction and instructional behavior modification as well as in session monitoring. The job models and the Session Monitoring Form both encompassed behaviors specifically addressed by the BARS in the present study.

The scale dimensions consisted of:

1. S\textsuperscript{D} presentation requiring attending by the student, a 1-second pause for student attending by the instructor before presenting a clear and correct S\textsuperscript{D} as described within the Distar lesson.

2. Prompting requiring timelines (not before 5 seconds and not after 10 seconds of no responding or immediately after a wrong response) and achievement of a target response as stipulated within the Distar lesson.

3. Consequation in three modes or dimensions with appropriate reinforcement within 1 second of a response, including descriptive praise and a tangible reward, such as a token, edible, or physical contact.

It should be noted that the scales developed for the present study were much more specific and limited in scope than the BARS developed in previous studies for psychometric evaluation and quarterly to annually general application. Actually, the scales in the present study were not intended for general performance evaluation but were intended: (a) to determine if BARS presentation in a controlled setting could directly affect performance in the above categories which were crucial to effective teaching; and (b) to evaluate that performance.
for intraclassroom management decisions. Because of the specific content of the scales and the value of the anchors, the present BARS may be even more objective than those seen in previous studies so the scales may be considered Behaviorally Anchored Rating Scales. But because of the limited scope of the scales and the method of their application in the present study, the BARS presented here could not suffice as a total BARS evaluation system.

Experimental Design

A simple multiple baseline design across behaviors was utilized with Subjects A, B, C, and D. A reversal was also conducted with Subject C. The orders of presentation of scales for each subject are presented: Subject A—consequation, prompting, and \( S^D \) presentation; Subject B—consequation, \( S^D \) presentation, and prompting; Subject C—consequation, prompting, and \( S^D \) presentation; and Subject D—prompting, \( S^D \) presentation, and consequation. The order of intervention was determined by deficits in the instructors' repertoires and reflected an attempt to provide the best instruction possible for the students.

Data Collection

A standard session monitoring form was used to record data about the instructors' teaching behaviors. As mentioned before, the observer (in this case, the experimente and present author) sat behind the students in a position facing the instructor. For each trial which began with each new instruction, a plus or a minus was marked to.
indicate whether or not (a) the student was attending; (b) there was a 1-second pause by the instructor for student attending; (c) a correct and clear $S^D$ was presented; (d) a correct response occurred and (e) if not, did a prompt occur; (f) if the prompt was not correct, was it late or did no target response occur; and (g) was the consequation correct (appropriate, immediate, descriptive, and inclusive of tangible rewards). Appropriateness of consequation included nonreinforcement of incorrect responses and reinforcement of some form for correct responses. Inappropriate reinforcement rendered consequation incorrect in all three dimensions for the respective trial.

BARS Presentation

Subjects A and B were exposed to the BARS immediately prior to (within 10 minutes of) the session following the one in which they were observed and rated. This was done to assess any antecedent effects of the BARS upon staff behaviors. Subjects C and D were exposed to the BARS immediately after (within 10 seconds of) the session in which they were observed and rated. This was done to assess any consequent effects of the BARS upon staff behaviors. A summary of experimenter-subject interaction is provided in Appendix E.

The BARS ratings were always stated in terms of lower limits set down within each scale. For example, if 85% of reinforcers occurred within 1 second of correct responses, this was stated as 70% as represented by the lower limit on the scale. The numerical rating always represented the anchor describing the lowest percentage value for each segment of the dimension. Similarly, if $S^D$'s were exact at less than
50% occurrence while attending, 1-second pauses, and clear $S^D$s occurred at 50%, the scale value of 2 was circled to represent "wrong $S^D$"; and then the remaining descriptive performance anchors were underlined to avail the subjects to the most specific feedback possible with this form.

**Reliability**

Reliability of the observer was assessed across several exposures to the same set of behavioral events since "assessing the stability of observer judgments is most easily accomplished if a constant behavioral event can be repeated, presented, and observed" (Johnston & Pennypacker, 1980, p. 195). Two videotaped sessions for each subject served as the repeated behavioral events. The tapings were initially made in the third sessions for Subjects B and D and in the fourth sessions for Subjects A and C. This was done to allow the students and instructors to become desensitized to the new system before introducing a new and intriguing stimulus and to allow the use of the only available videotape with all subjects. The second tapings were made in Session 9 with Subjects B and D and in Session 10 with Subjects A and C. This was done at this point simply to recalibrate the observer and to allow the use of the one tape. The videotape equipment was placed to the right side of all of the students to allow a view of each of them. Reliability was computed by dividing all the agreements by all the agreements plus disagreements for the several observations of each individual session.
RESULTS

The results of the present study are presented graphically in Figures 1 – 4. Three intervention elements are displayed in each multiple baseline graph. $S^D$ presentation is displayed in terms of student attending (represented by large solid dots connected by solid lines), a 1-second pause for attending by the instructor (represented by open circles connected by dashed lines), and correct $S^D$'s (represented by open squares connected by dotted lines).

The prompting element simply consists of the percentage of correct prompts by each subject. An arrow points down from the upper limit of this element in some cases to indicate that one incorrect prompt occurred in less than 10 opportunities to prompt. This is important to note because this affects more greatly the variability in the percentages here than in the other elements, since in the latter, there were always at least 10 opportunities for responses, and in most cases, 20 or more. And, there was no control of opportunities to prompt, since this depended on incorrect responses by the students.

Consequation is displayed as percentage correct in three dimensions of behavior (praise, description of the behavior, and a tangible reward) and in two dimensions (any two of the previous dimensions). It refers to reinforcement of correct behaviors with nonreinforcement of correct behaviors or reinforcement of incorrect behaviors counting as incorrect in all three dimensions. In the graph for Subject A, an arrow points down from the upper limit of consequation at Session 10 to indicate that the instructor changed the procedure to a VR2.
Figure 1: Instructor Consequation, Prompting, and SD Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject A.
Figure 2: Instructor Consequation, Prompting, and $S^D$ Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject B.
Figure 3: Instructor Consequation, Prompting, and SD Presentation and Their Effects upon Student Responses as a Function of Consequent BAR Presentation for Subject C.
Figure 4: Instructor Consequation, Prompting, and ZD Presentation and Their Effects upon Student Responses as a Function of Antecedent BARS Presentation for Subject D.
schedule of token consequation to fade it out as was customary in many procedures in her classroom. Here the scale on consequation served as feedback to indicate to the instructor whether token or tangible consequation was occurring at, above, or below 50%. This simply indicates that the criterion was 50% in the three dimension component, while the experimenter still sought perfection in the remaining two dimensions.

Student responses are also displayed with each subject's data to demonstrate any indirect effect of the BARS on these through the subjects' instructional behaviors. The percentage of correct responses is displayed by open circles connected by dashed lines, and the percentage of correct plus corrected responses (total percentage of evoked responses) is displayed by large solid dots connected by solid lines.

With Subject A, where BARS was used as an antecedent, its presentation apparently had an initially deleterious effect upon her consequation of student behaviors. Consequation in baseline ranged from 77% to 84% in three dimensions and from 93% to 100% in two dimensions. With the advent of antecedent BARS presentation, consequation in three dimensions fell to 64% and in two dimensions to 82%. Over the five following sessions, these seem to recover somewhat. The data begin to follow a predictable pattern at Session 10, where the VR2 schedule with tokens was implemented. The VR2 token schedule seems to have negatively affected the other two dimensions of reinforcement in Session, but reinforcement in two dimensions shows gradual improvement thereafter. Reinforcement in three dimensions after
Session 10 stayed within 10 percentage points of criterion in five out of eight sessions. In Session 17, Subject A wanted to discontinue the use of tokens, so the experimenter reminded her to use physical contact instead. Here consequation in three dimensions fell to 33%. Reinforcement in three dimensions never deviated more than 20 percentage points from the 50% criterion.

Percentage of correct prompting with Subject A in the initial baseline hovered around 50%, with an overall increase in intervention to about 80%. The graph spikes to 100% four times during the intervention, about 35 percentage points higher than the highest percentage during baseline. In SD presentation, the experimenter waited until the data showed some variability before intervening with the BARS there, since all the components of SD presentation (attending, 1-second pause, and correct S^D) were hovering at or around ceiling. There is an upward trend in student attending and in the 1-second pause from 74% to 100% in the BARS intervention. However, due to the end of the school year, no further data were obtainable; and given more data, perhaps a more definite trend would have appeared.

Student responses throughout the study appear to improve in terms of correct responses and responses corrected by prompts. The latter necessarily co-varied to some extent with the percentage of correct prompts by the instructor. But the percentage of correct student responses also appears to co-vary with the percentage of correct prompts by Subject A. This is particularly evident in Sessions 9 through 14, where in every instance except Session 12, the percentage of correct student responses shows a direct relationship with the
percentage of correct prompts by the instructor. Correct student responses increase from 50% in Session 9 to 73% and 90% in Sessions 10 and 11 respectively with the advent of a VR2 schedule of token reinforcement. But after Session 11, student response rates fell to a lower but more consistent rate of about 65% for the remainder of the study.

Consequation of student responses by Subject B shows a substantial amount of variability in baseline. The range of percentage points is from 20 to 87 in three dimensions of consequation and 68 to 100 in two dimensions of consequation with no apparent general trend. With the introduction of antecedent BARS presentation, this variability is reduced with a substantial increase in correct consequation in two and three dimensions. The range of scores in BARS intervention is 60% to 100% in three dimensions of consequation and 77% to 100% in two dimensions of consequation.

$s^D$ presentation in all three categories (student attending, 1-second pause, and correct $s^D$) with Subject B shows some initial variability but then approaches ceiling with some variability while still in baseline. Intervention was required at Session 11 because of the nearing of the close of the school year. However, even with a lack of room for improvement, the data do show some increase and narrowing of variability in percentage points for attending and correct $s^D$s and for the 1-second pause except for Session 13.

The baseline data for Subject B in prompting may be deceiving. In Sessions 5, 7, 8, and 11, only one incorrect prompt occurred. Because of this, the data could have been very much different given
more opportunities to prompt. Conceivably, less variability could have occurred if only a few more correct prompts were delivered in Sessions 5 and 8 and even in Session 6, where only two opportunities for prompting occurred but which yielded 0%. On the other hand, given more incorrect prompts in Sessions 7 and 11, the data again may have looked very different. Therefore, due to the instability and variability in the baseline prompting data, the experimenter waited as long as possible to implement antecedent BARS presentation for instructor prompting. Upon BARS intervention, no narrowing of variability from the last four baseline data days is evident.

The rate of correct student responses shows little relationship with the percentage of correct prompts by Subject B. However, it does show a direct relationship with student attending in Sessions 1 through 8 and some co-variation in the remainder of the sessions. In addition, correct student responses co-vary with the 1-second pause throughout the study except in Sessions 1, 6, 10, and 13. And they co-vary with conseuation in three dimensions in Sessions 2 through 9 and in Session 14.

With Subject C, one fact is readily apparent. That is, that no reversal occurred in any element. And in every element, the same general statement can be made about the effects of consequent BARS presentation. There was moderate variability in baseline with some apparent improvement in instructor responses given the introduction of the BARS. There is still further improvement in the reversal and in BARS reintroduction respectively. Percentage of correct student responses does not apparently co-vary with any element or any segment.
thereof. And there is no general increase over time in the percentage of correct student responses.

The baseline prompting data for Subject D show a downward trend from about 80% to about 40%. With the introduction of consequent BARS presentation, a substantial increase in percentage of correct prompts is observed from Session 5 to 7, with a gradual decrease from Session 7 to 10. Again, a substantial increase from 50% to 100% occurs in the remaining four sessions.

$S^D$ presentation by Subject D shows a moderate amount of variability in baseline, especially for attending and the 1-second pause. With the introduction of the BARS, there is both a reduction in variability of and a general increase in the percentage of correct $S^D$'s, student attending, and the 1-second pause. The percentage of correct consequation in two and three dimensions in baseline is very high and shows little variability. With the introduction of BARS presentation, however, consequation in three dimensions falls drastically, while consequation in three dimensions remains very high. Corrected student responses appear to co-vary with consequation in two dimensions, and there appears to be a slight improvement in both categories of student responses over the course of the study.

The reliability data for the observer were calculated as the total number of agreements over the total number of agreements plus disagreements in three repeated observations and data recordings of two different sets of stimuli (two different videotapes of actual experimental sessions for each subject) between subjects over all the dimensions. The reliability data for Viewing A are: Subject A, 92%;
Subject B, 98%; Subject C, 94%; and Subject D, 89%. The reliability data for viewing B are: Subject A, 95%; Subject B, 91%; Subject C, 97%; and Subject D, 90%.
DISCUSSION

It is apparent after a general overview of the results of this study that the introduction of the BARS had a slight functional relationship with an increase in the percentage of correct instructor responses. This is especially true in the case of Subject A, where consequation in three dimensions conforms close to the 50% criterion line when nowhere earlier in her case did this occur. BARS effect on instructor responses was also apparent with Subject B in her consequation of student behaviors and to a lesser extent in her SD presentation. With Subject C, the BARS appear to have had a slight effect upon instructor responses in all the elements; but the continued gradual increase in these responses during reversal is puzzling. And finally, except for consequation, Subject D's instructional responses appear to improve with the introduction of BARS presentation.

Before an analysis of the success of BARS is made, a discussion of its shortcomings is in order. The first major apparent failure of BARS was in its effects upon consequation for Subject D. The variable that most dramatically occurs along with the dip in consequation is a break in sessions of 5 data days between Sessions 9 and 10 due to a long Memorial Day closing of the facility where this study was conducted. Another variable that occurred here was that Subject D was transferred to a different classroom after that break than the one she worked in before it. Due to this transfer, her schedule conflicted with the schedule normally maintained for this study; and she engaged in a great deal of behavior to enable herself to participate in this
study. These variables may have had some effect on her performance in the instructional sessions. Whether the cause of the drop in performance was due to these or other unsuspected variables, it apparently also had a negative effect on prompting by Subject D, although it was not such a profound effect. The possibility also exists that the observer was lenient or lax in the recording of consequation in three dimensions during baseline and then became more observant of reinforcement with the introduction of BARS for instructor consequation. But this is not consistent with the data for the other subjects, nor with the reliability data.

Another problem with the consistency of BARS effects on instructor behaviors is the lack of a reversal in any element for Subject C. There are several possible explanations. First, the BARS may have had little or no effect on her behavior with the improvement in her responses being due to simple practice effects. Conceivably, reinforcement for her responses could have been programming by growing familiarity with the materials or by success with student compliance as a function of following the written Distar materials. Another possibility is that the feedback contained in the BARS evoked rule-governed behavior and that this persisted from the introduction of BARS. This latter explanation seems the more parsimonious since the initial introduction of BARS in consequation and in $S^D$ presentation very quickly reduced variability in the data and improved the instructor's responses. This is also true to a lesser extent with prompting. In other words, the multiple baseline appears to have shown an effect of BARS introduction.

The major effect of BARS in this study seems to have been due to
its feedback characteristics. The scales did not contain particularly positive or negative labels to serve as reinforcers or punishers. The subjects did, however, voice displeasure with low rankings and pleasure with high rankings. The observer who presented the scales would, upon the achievement of a 10 (the highest rating), show the scale or scales to the subject and announce, "You're a 10." The subjects all seemed to enjoy this. But other than this, there was no inherent reinforcing or punishing stimulus associated with the scales. The subjects all volunteered the information that the scales served two functions for them. First, they showed the instructor exactly what she did in conjunction with a discussion of her performance with the observer. Secondly, they showed her exactly what it was she had to do to attain the highest ranking. Finally, the observer was acquainted socially with all the subjects. This may have rendered low scores socially punishing and high scores socially reinforcing.

Consistent with the findings of Krumhus and Malott (1980), there was no significant difference between the effects of BARS presentation as an antecedent and as a consequence. However, the data for antecedent BARS presentation follow a generally more predictable pattern than the data for consequent BARS presentation. If the effects of BARS are due mainly to their feedback characteristics, then that feedback may be more effective in evoking rule-governed behavior when it is presented immediately prior to the targeted behaviors, while its effects in evoking rule-governed behavior may diminish over time. However, this is an issue that should be deliberated in the feedback literature. The data in the present study provide nothing but
speculation as to the differences between antecedent and consequent BARS presentation.

The validation for the study of the effects of using BARS as a training tool is evident in the data for student responses. The first question personnel managers might have about BARS is "Will it improve performance?" That has been shown to have occurred to some extent. The second question is "Will it have a positive effect on the receiving system, the environment, or the clientelle?" In other words, will it result in real organizational improvement and benefits? Positive effects in the receiving system are demonstrated to a moderate but fairly consistent extent with the improvement in student responses. In Distar Math and Language programs, an overall success rate of 75% correct student responses is the goal. In the case of Subjects A and D, corrected student responses went from under 75% to 100%; and they went from under 90% to 100% for Subjects B and C. A 25% error rate is supposed to allow for a slight amount of programmed failure to give the students some information as to what not to do and to make reinforcers more effective. But in this study, the high response rates served to validate the use of BARS as a training tool.

The present study differs experimentally from previous studies in BARS application since those, almost exclusively, have dealt with psychometric comparisons of BARS with other rating formats and have not dealt with effects of BARS on behavior or other organizational results. Ivancevich (1980b) reported an improvement in scheduling performance by professional employees, but this was not quantified. This study was an attempt on a basic and empirical level to determine
whether BARS would have any true effect upon performance. The importance of this study may not lie in the results of the experiment (or the lack thereof), but it may well lie in the precedent it has set in seeking and establishing behavioral goals for and effects of BARS application.

Landy and Barnes (1979) have set down specific points of logic in using BARS. The people who use the scales have to some extent in most cases been its developers as well. More information is provided with behavioral anchors than with simple heading descriptors. These anchors have similar meanings to all raters, and the performance dimensions in BARS are easily distinguishable from one another.

And a rising need for accurate performance appraisals makes BARS perhaps the best choice of several evils since, as Churchill said of democracy, "It's a terrible system but it's the best there is."

Keaveny and McGann (1975) point out that accurate performance appraisals allow for proper reward and promotion for effective and high performing personnel. This would be a break for employers who are constantly seeking better qualified persons to fill important positions. On the other end of the spectrum, white collar employees are disputing discharges more frequently; and an accurate appraisal of poor performance would fairly support such discharges. So in terms of real organizational costs, accurate and effective appraisals of performance are a grave necessity.

The variation of the scales used in the present study from the typical BARS approaches a similarity to the Behavior Observation Scale, or BOS. The intent of the BOS is to equalize the value of
respective levels of performance across behavior items (Latham, Fay, & Saari, 1979). The BOS consists of performance dimensions named "criteria." with subsets of "behavioral items." The criteria and items are generated in a manner similar to the development of performance dimensions and behavioral anchors for a BARS. The behavioral items in a BOS number from 5 to 10 in each of 5 to 10 criteria. The items are underscored by groups of observational frequencies (i.e., 0 - 20%, 21 - 40%, 41 - 60%, 61 - 80%, and 81 - 100% frequency of observation of the behavioral item), the appropriate category of which is checked at the time of the performance evaluation (Kane & Bernardin, 1982).

As with every performance evaluation tool reviewed by the present author, problems with the BOS have been presented. Kane and Bernardin (1982) claim that the goal of the BOS, which is to equalize respective levels of frequency, is fallible in that equality of frequency of observation does not necessitate equality of the weight of the frequency levels between items. For instance, they point out that equal frequencies of obtaining homicide arrest warrants and being cleared by the Internal Review Board for the use of fatal force by a particular police officer is unacceptable. Eighty percent frequency of obtaining homicide arrest warrants may be superior, but being cleared by the Interval Review Board 80% of the time is horrendous. Latham, Saari, and Fay (1980) claim that this equality of frequency levels does exist but no definitive resolutions have been achieved and the debate has become heated (Bernardin & Kane, 1980; Latham, Saari, & Fay, 1980).
In review, the BARS system consists of performance dimensions underscored by a scale which represents poor to excellent performance with behavioral anchors attached to the scale values describing examples of poor to excellent performance. The scales used in the present study consist of performance dimensions underscored by scales representing poor to excellent performance with behavioral anchors consisting of frequencies of observation of sets of behaviors attached to the scale values to describe poor to excellent performance. It is these frequencies of performance which resemble the frequencies of observation of behaviors in the BOS. While these frequencies do resemble, and indeed serve a similar purpose as, the BOS, there are important differences to make clear. First, the frequencies provided within the scales in the present study are a departure from a typical BARS which consist of adjectival rather than numerical descriptions of poor to excellent performance comprising a performance dimension. This departure was utilized to maximize the objective observation and evaluation of subject behavior and was conceived to make an empirical assessment of the BARS as a training tool rather than to create a new performance evaluation tool. Second, the respective frequency levels of performance are not and were not intended to be parallel across performance dimensions; and the observational frequency levels must be parallel across behavior items to qualify as a BOS.

Despite the fact that the present study concentrated on assessing the BARS as a training tool, it does not mean that it may not provide important information applicable to the BOS and other evaluation tools. As has been pointed out earlier, it may well be the case that the
method of obtaining the information pertinent to performance evaluation, the information actually used, and the way it is used (i.e., will any change in the environment occur as a result of the evaluation) are more important than the name of the tool used or the way the information is displayed.

Peterson (1982) provides some suggestions for the use of terminology in feedback studies. He claims that the term "feedback" has become "professional slang" and that it should not be used as a description of the interventional procedure. He suggests that authors in organizational behavior management literature do not sufficiently define the specific feedback procedures used and that these specific descriptions must be provided to interpret and analyze the procedures and their results. This, he says, will facilitate the analysis of the effects of feedback and that this is of paramount importance in order to move away from unfounded assumptions in the literature about the effects feedback has on performance so that feedback effects can be assessed. When descriptions of specific feedback procedures are not provided in the literature, the behavioral effect of feedback (i.e., reinforcement, punishment, stimulus control) cannot be assumed; and the research cannot be interpreted in any meaningful way. Specifying experimental procedures and their results should be, in this author's opinion, the focus of subsequent studies in using feedback to alter behavior and improve performance. These points can also be applied to studies involving rating scale evaluations of performance since adequate descriptions of the way in which the rating scales are used has frequently not been specified in performance evaluation studies.
Remedies to these problems with feedback and evaluation studies were sought in the present study by providing the scales used and detailed descriptions of the procedures.

More studies like the present one should be conducted in the future to determine the extent of the effect BARS can have on an organization. The author would note that it has been encouraging to see so many studies devoted to the development of clear and objective statements about organizational performance. As Drucker (1973) has said:

An employer has no business with a man's personality. Employment is a specific contract calling for specific performance, and for nothing else. Any attempt of an employer to go beyond this is usurpation. It is immoral as well as illegal intrusion into privacy. It is abuse of power. An employee owes no "loyalty"—he owes performance and nothing else. . . . Management and manager development should concern themselves with changes in behavior likely to make a man more effective. (p. 424-425)

It is the conclusion of the author that the BARS may serve as an effective training tool. This may be accomplished by using the scale itself as a feedback mechanism as was done in this study, or by using the information gained in traditional BARS development procedures to develop a programmed instruction or other behavior-based training program. But before this can be a certainty, more evaluations of BARS as a training tool must be made. The author suggests a move away from psychometric evaluations of BARS unless new developments are manifest in BARS and other performance evaluation procedures. As Bernardin and Kane (1980) point out, "It is our view that another round of empirical comparisons between formats that do not depart appreciably from previously tested methods is unwarranted" (p. 810). A behavioral evaluation of
BARS by experienced developers would be paramount. With the advent of BARS that are developed through the more traditional means yielding psychometrically sound scales, more profound training effects of BARS than in this study may become apparent.
### APPENDIX A: $S^D$ PRESENTATION

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>90%—Attending&lt;br&gt;90%—$S^D$, signal exact&lt;br&gt;90%—Clear presentation</td>
</tr>
<tr>
<td>9</td>
<td>90%—Attending&lt;br&gt;90%—$S^D$, signal exact&lt;br&gt;90%—Clear presentation&lt;br&gt;70%—1-sec. pause</td>
</tr>
<tr>
<td>8</td>
<td>90%—Attending&lt;br&gt;70%—$S^D$, signal exact&lt;br&gt;70%—Clear presentation&lt;br&gt;70%—1-sec. pause</td>
</tr>
<tr>
<td>7</td>
<td>90%—Attending&lt;br&gt;70%—$S^D$, signal exact&lt;br&gt;70%—Clear presentation&lt;br&gt;70%—1-sec. pause</td>
</tr>
<tr>
<td>6</td>
<td>70%—Attending&lt;br&gt;70%—$S^D$, signal exact&lt;br&gt;70%—Clear presentation&lt;br&gt;50%—1-sec. pause&lt;br&gt;50%—Unclear $S^D$&lt;br&gt;50%—$S^D$, signal inexact</td>
</tr>
<tr>
<td>5</td>
<td>70%—Attending&lt;br&gt;50%—1-sec. pause&lt;br&gt;50%—$S^D$, signal inexact&lt;br&gt;50%—$S^D$ unclear</td>
</tr>
<tr>
<td>4</td>
<td>No $S^D$</td>
</tr>
<tr>
<td>3</td>
<td>Wrong $S^D$</td>
</tr>
</tbody>
</table>
| 2     | Wrong $S^D$
No attending |
| 1     | No attending |
| 0     | No attending |

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
APPENDIX B: PROMPTING

90%—After 3 sec. of no responding
90%—After wrong response
90%—Target response achieved

70%—After 3 sec. of no responding
70%—After wrong response
90%—Target response achieved

50%—Sometime after 3 sec. of no responding
50%—More than 3 sec. after wrong response
70%—Target response achieved

50%—Sometime after 3 sec. of no responding
50%—More than 3 sec. after wrong response
50%—Target response achieved

70%—No target response
50%—Incorrect prompt

50%—No prompt after no or wrong response

0
1
2
3
4
5
6
7
8
9
10
APPENDIX C: CONSEQUATION

90%—S^r+ w/in 1 sec. of response
90%—Tangible S^r+, description, immediate praise (three dimensions)

70%—S^r+ w/in 1 sec. of response
90%—S^r+ in three dimensions

50%—S^r+ up to 5 sec. after response
90%—S^r+ in three dimensions

50%—S^r+ up to 5 sec. after response
90%—S^r+ in two dimensions

30%—No S^r+ for response

30%—S^r+ for wrong response

50%—S^r+ for wrong response

50%—no S^r+ for response

40%—S^r+ for wrong response
APPENDIX D: SESSION MONITORING FORM

<table>
<thead>
<tr>
<th>Monitor date:</th>
<th>Is the procedure affecting behavior change in the desired direction? YES NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days on this step/phase:</td>
<td>_________________________________________________</td>
</tr>
<tr>
<td>General comments:</td>
<td>_________________________________________________</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tutor signature:</th>
<th>Monitor date:</th>
<th>Is the procedure affecting behavior change in the desired direction? YES NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days on this step/phase:</td>
<td>_________________________________________________</td>
<td></td>
</tr>
<tr>
<td>General comments:</td>
<td>_________________________________________________</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tutor signature:</th>
<th>Monitor date:</th>
<th>Is the procedure affecting behavior change in the desired direction? YES NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days on this step/phase:</td>
<td>_________________________________________________</td>
<td></td>
</tr>
<tr>
<td>General comments:</td>
<td>_________________________________________________</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX E: INTERACTIONS BETWEEN OBSERVER AND SUBJECT

When the BARS were presented as antecedents and as consequences, the experimenter interacted with the subjects to explain exactly what the scale feedback represented. Examples of this interaction follow.

The experimenter interacted with each subject individually. First, he would present the scale(s) and simply recite the information contained in the scale. Then, he would relate that information to the previous day's performance on each scale. In other words, "You scored a 4 today, and that is worse than yesterday," or "You scored a 9 today, and that is better than yesterday." This was done to give the subjects some idea of the progress of their performance. After these daily interactions, the experimenter asked the subjects if there were any questions about the scale, their performance, or the experiment. Almost exclusively, they would ask for an analysis of their performance, such as, "How should I present this new task?" or "What should I do about a child's behavior problem?" Or they would ask what they should do specifically to improve their rating. Occasionally, they would make suggestions for amendment of the written procedure. And every subject asked at least once how the experiment in general was going, but this was just out of curiosity about how an experiment is run.

Regarding questions asked about how to deal with the students, because the student's best interest was of greatest importance, the observer would do whatever was necessary to help the subject deal with the student, especially in the case of the behavior problem. If the question dealt with the specific Distar procedures involved, the subject was first directed to read the manual's instructions; and next, if necessary, the subject was directed to read over the rating scales.
since these included all the information necessary to conducting an
instructional session within the facility. And regarding questions
about the research project, they were answered as honestly as possible
since this was a policy of the experimenter.
BIBLIOGRAPHY


Borman, W. C., & Vallon, W. R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology, 1974, 50, 197-201.


DeCotiis, T. A. An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977, 19, 247-266.


