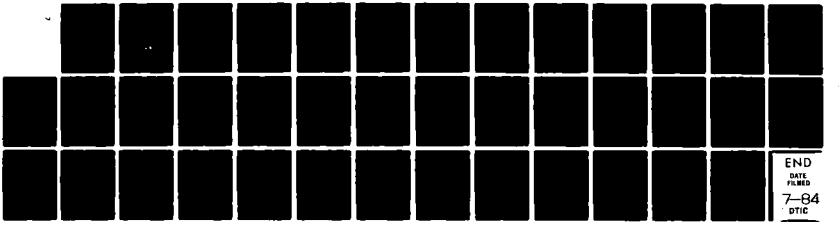
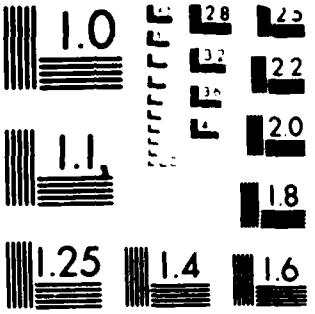
AD-A142 009 THE DEVELOPMENT OF TRAINING PROGRAMS TO INCREASE 1/1
ACCURACY WITH DIFFERENT..(U) MICHIGAN STATE UNIV EAST
LANSING DEPT OF PSYCHOLOGY E D PULAKOS MAY 84 TR-84-2
UNCLASSIFIED N00014-83-K-0756 F/G 5/9 NL



END
DATE FILMED
7-84
DTIC



MICHIGAN STATE UNIVERSITY

Industrial/Organizational Psychology
and
Organizational Behavior

The Development of Training Programs to Increase Accuracy with Different Rating Formats

by

Elaine D. Pulakos

Michigan State University

Michigan State University
East Lansing, Michigan 48824

The Development of Training Programs to Increase Accuracy with Different Rating Formats

by

Elaine D. Pulakos

Michigan State University

Prepared for
Office of Naval Research
Organizational Effectiveness Research Programs
Code 4420E

Grant No. N00014-83-K-0756
NR170-961

.Technical Report 84-2
Department of Psychology
and
Department of Management
Michigan State University

A

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 84-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Development of Training Programs to Increase Accuracy with Different Rating Formats		5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG. REPORT NUMBER 2004
7. AUTHOR(s) Elaine D. Pulakos		8. CONTRACT OR GRANT NUMBER(s) N00014-83-K-0756
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology Michigan State University East Lansing, MI 48824-1117		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR170-961
11. CONTROLLING OFFICE NAME AND ADDRESS Organizational Effectiveness Research Programs Office of Naval Research (Code 4420E) Arlington, VA 22217		12. REPORT DATE May, 1984 13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Performance Appraisal, Accuracy, Rater Training, Rating Format

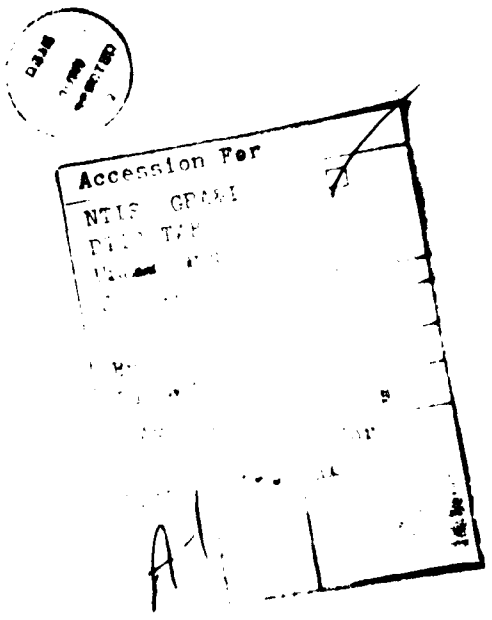
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The research investigated potential interactive effects of format and training on rating accuracy. Training was developed in accordance with the cognitive demands placed on raters by each of two rating formats: behaviorally anchored rating scales (BARS) and behavioral observation scales (BOS). In order to investigate format x training interactions, the effects of each training procedure were assessed for both types of scales. In addition to increasing accuracy, congruent training (BARS training with BARS and BOS training with BOS) was hypothesized to increase interrater agreement. One hundred and forty-four

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

subjects were randomly assigned to 1 of 6 cells defined by combinations of training (BARS, BOS, control) and format (BARS and BOS). Analyses of variance revealed that for those using a BARS format, ratings from the BARS training group were more accurate than ratings from the BOS training group, which, in turn, were more accurate than ratings from the control group. When BOS were used to rate performance, only BOS training led to higher rater accuracy. Interrater agreement was greater in congruent training groups. Implications, limitations, and future research directions are discussed.



The Development of Training Programs to Increase Accuracy with Different Rating Formats

Researchers in the performance appraisal area have generally adopted one of two strategies to increase the reliability and validity of performance ratings: (1) improving the rating formats or (2) training raters. Unfortunately, the results of many format comparison studies indicate that scale modification has not been a particularly useful strategy for improving performance ratings (Bernardin, Alvares, & Cranny, 1978; Bernardin & Kane, 1981; Landy & Farr, 1980; Schwab, Heneman, & DeCotiis, 1975). Rater training programs have been successful in reducing common rating errors (Bernardin, 1978; Latham, Wexley, & Purcell, 1975). However, error training has been found to have virtually no effect on rating accuracy (Borman, 1975, 1979; Pulakos, 1984).

Several authors have recently suggested that accuracy might be increased by training raters to use a common frame-of-reference for observing, interpreting, and judging ratee performance (Bernardin & Buckley, 1981; Borman, 1979, 1983). Preliminary support for this notion has been provided by McIntyre, Smith, and Hassett (1984) and Pulakos (1984) who attempted to impose on trainees a set of standards for evaluating performance based on the dimensional structures of the rating scales being used. The rationale underlying these training procedures was similar to that associated with behaviorally anchored rating scales (BARS; Smith & Kendall, 1963); referent anchors were provided to facilitate agreement in evaluating

the effectiveness of ratee behavior. In both studies, training was shown to significantly increase rating accuracy.

Given that these "frame-of-reference" training programs presumably oriented trainees to use common evaluative standards for appraising performance, they are intuitively very appealing. However, it is not entirely clear whether or not this type of training would be equally effective with all types of rating scales. After all, such a global strategy does not take into account the particular and potentially different cognitive demands that are placed on raters by various rating formats. The purpose of the present research was to investigate potential interactive effects of format and training on rating accuracy. First, training programs were specifically developed so as to be congruent with the different cognitive demands placed on raters by two popular rating scale formats: BARS and behavioral observation scales (BOS; Latham & Wexley, 1977; 1981). Training x format interactions were then assessed by evaluating the effects of each general training procedure on ratings that were made using both formats.

Rating Scales and Training

BARS. When using a BARS format, raters evaluate ratees on several (usually five to ten) job performance dimensions. Each dimension, usually presented on a single page, is defined by both a general description and a number of scaled behavioral anchors ranging from excellent to poor performance. The rating task involves selecting a level of evaluation that best describes the

effectiveness of ratee performance on each dimension.

Given that evaluation is the focus in appraising performance with BARS, the previously discussed "frame-of-reference" training should be a reasonable strategy for increasing accuracy. Through providing an understanding of the dimensional system itself and the effectiveness of behaviors that attend upon it, this type of training should promote the use of uniform standards for judging performance.

BOS. Although BOS and BARS are structurally similar, (i. e., they each contain several performance dimensions which are further defined by examples of specific employee behaviors), different components of the rating task are emphasized with each format (Murphy, Martin, & Garcia, 1982). Whereas BARS focus on evaluation, BOS are primarily concerned with observation and require that raters report the frequency with which a number of critical ratee behaviors have occurred. Because complex evaluative judgments are not required in using BOS, training individuals to use common standards for judging performance effectiveness may not be the optimal strategy for increasing rating accuracy. Rather, given the demands placed on raters by BOS, it seems that accuracy would be better facilitated by sharpening raters' observational skills (so that critical behaviors are recognized quickly) and by providing them with a strategy to aid their recall of how often relevant ratee behaviors occurred.

In addition to evaluating training effects in general, also of

interest here was the assessment of any potential differences in training program effectiveness for different job performance dimensions. Finally, it was hypothesized that when each training program was used with the format for which it was intended, trainees should come to consider ratee performance in similar ways (i. e., appropriate, albeit different, frames-of-reference should be developed for use with each type of scale). Accordingly, ratings should be more accurate and interrater agreement should be higher within congruent training groups (BARS TRNG with BARS and BOS TRNG with BOS) than within incongruent or no training conditions.

Method

Subjects

Participants in the study were 144 undergraduate students (80 females and 64 males) enrolled in an introductory psychology course. Their mean age was 19.42 years, and 83 of the students reported having previous experience with performance appraisal. Students were randomly assigned to one of six experimental groups (N = 24 per group).

Design and Procedure

The research was explained to potential subjects by informing them that the study involved using performance appraisals to evaluate managers seen on videotapes talking with a problem subordinate. Extra credit points were given to students who agreed to participate.

Three training conditions were created. These were BARS congruent training, BOS congruent training, and a control training program labeled here "control training." The latter was created

in order to keep the laboratory time consistent for all three groups. Crossed with all three training conditions were two rating formats (BARS and BOS) and five performance dimensions. The dimensions were nested within each training x rating scale format cell.

Videotapes and Behaviorally Anchored Rating Scales

Subjects viewed 5- to 9-minute videotapes of eight managers talking with a problem subordinate. Videotaped performances were used because they enabled the calculation of true scores, thereby allowing an assessment of rating accuracy. Further, the videotapes were carefully developed so as to ensure that the performances represented a variety of effectiveness levels on different rating dimensions. Specific details regarding the development of the tapes, the BARS, and the procedure used to generate true scores for the BARS can be found in Borman (1977).

Ratings of each manager's performance were made on the following performance dimensions: (1) Structuring and Controlling the Interview; Establishing and Maintaining Rapport; (3) Resolving Conflict; (4) Motivating the Subordinate; and (5) Developing the Subordinate. Each dimension was defined by an overall statement and contained seven, scaled behavioral anchors describing different effectiveness levels.

Behavioral Observation Scales

Critical incidents originally collected in developing the BARS were used to develop the BOS. Specifically, BOS items were based on critical effective and ineffective manager behaviors that had been

reliably retranslated into a particular dimension and effectiveness level. In order to ensure a high degree of correspondence between observed and scaled behaviors, two graduate students carefully reviewed the videotapes and rewrote items as necessary to match more directly with the behaviors actually exhibited by the managers. These procedures resulted in a total of twenty-four critical incidents, each representing one of the five performance dimensions. For each item, raters were asked to indicate, on a Likert-type scale ranging from 0 to 3+, the number of times each manager exhibited the behavior.

Ten expert raters were selected to evaluate the effectiveness of the managers using the BOS. Similar to the way in which Borman (1977) generated the BARS true scores, expert raters were given scripts of the videotapes and the rating scales, and they were asked to study these prior to rating the managers. True scores were then determined for each item if at least eight of the ten raters agreed on the number of times (from 0 to 3+) each manager exhibited the behavior. Interrater agreement was considerable for the BOS: there was 100% agreement on 76% of the items, 90% agreement on 18% of the items, and 80% agreement on the remaining 6% of the items. Further, the correlations between the BARS and BOS true scores were also high, ranging (by dimension) from .91 to .97 with a mean $\underline{r}$ of .94.

Rater Accuracy Training Programs

BARS Training. Pulakos (1984) developed a training program to increase performance appraisal accuracy. This training focused on providing raters with a common set of standards for evaluating

ratee performance. This was accomplished by using a behavioral rating instrument (i. e., BARS) as a training tool along with focusing rater attention to the job performance dimensions and examples of what types of behaviors constituted various effectiveness levels within each. The general strategy used by Pulakos was employed here to train raters who would then use either a BARS or a BOS format to evaluate the videotaped managers.

Trainees were first given a lecture on the multidimensionality of jobs and the need to pay close attention to ratee behavior in terms of these dimensions. The actual rating scales that would be used to evaluate the managers were then distributed to participants (i. e., one group was given BARS and the other was given BOS) along with a separate list of the dimensions and their definitions. After discussing the meaning of each dimension, the trainer presented additional examples of specific behavioral incidents that corresponded to it. For those who would be rating with BARS, the trainer discussed the behaviors in terms of what might be expected of a manager who should be rated a "7" versus a "5" versus a "2" etc. For those who would be rating with BOS, the trainer discussed each of the behavioral incidents in terms of its general effectiveness level (i. e., high, average, or low) on the dimension.

Because the BOS group knew they would be rating the frequency with which particular behaviors occurred, it was necessary to provide them with an explanation that would legitimize their training program. Thus, prior to discussing the dimensions, the trainer explained that a

prerequisite to making accurate ratings was understanding the dimensional system of the scales, and such an understanding was facilitated by discussing the effectiveness of various behaviors that were representative of each performance category. Interviews with pilot test subjects indicated that this explanation seemed reasonable and did not foster any questions concerning the appropriateness of the training content for their particular rating task. Further, no questions were raised during the actual experimental sessions that suggested students might be skeptical about the training.

In both groups, subjects practiced using their respective rating scales by rating two of the eight videotaped managers. After viewing each tape, the group(s) discussed their ratings and received feedback on their accuracy. Those who were rating with BARS received evaluative feedback (i. e., which scale value was the correct rating for each dimension), while those who were rating with BOS received frequency feedback (i. e., how often each critical behavior occurred). Each training session lasted approximately one and one-half hours. Because two of the tapes were used for training, the results that follow are based on the remaining six manager performances.

BOS Training. In order to rate accurately using a BOS format, two activities seemed most critical. The first was that raters attend to and recognize critical behaviors that were exhibited by ratees. Second, because raters would be required recall frequency data, it seemed important that active attempts be made to keep a mental count of relevant behaviors. Further, congruent with the

BOS philosophy (Latham & Wexley, 1981), raters were encouraged to observe, rather than evaluate the effectiveness of, critical ratee behaviors. Those who received BOS training were first lectured on the importance of attending to relevant ratee behaviors (as opposed to traits) and on the difference between observation and evaluation. Participants were then given lists of the specific critical behaviors that corresponded to each rating dimension. In order to minimize the possibility that raters would form general impressions of effective versus ineffective performers (hence providing them with evaluative prototypes to which ratees could be matched), no mention was made of the degree to which the behaviors were characteristic of good or poor performance. Further, the behaviors were randomly ordered within each dimension as opposed to being listed, for example, in favorable and unfavorable subgroups.

Trainees were informed that their next task was to memorize the behaviors that appeared within each dimension. Specifically, they were told to read over the behaviors and to rehearse mentally ones that corresponded to each dimension. Without referring back to the scales, raters were asked to write down the dimension titles and the behaviors that fell within them. This task was repeated twice and subsequent to each trial, subjects corrected their responses by consulting the list of behaviors. The purpose of this exercise was to sharpen raters' observational skills by teaching them what particular behaviors they should recognize as important when observing the managers.

Trainees were also told that since their goal was to count behaviors, use of a mental checklist while viewing the tapes should facilitate their rating task. It was suggested that they view the tapes with the following in mind: Is this (behavior the manager is exhibiting) one of the critical behaviors I have memorized? If not, ignore it. If yes, has this manager exhibited the behavior before? If no, remember it occured once. If yes, how many times? Add N + 1.

As was necessary when BARS training was used with a BOS format, an explanation was provided to the BOS training/BARS format group that would legitimize their training content. Specifically, those rating with BARS were told that raters often make immediate evaluative judgme ts of ratees that are based on far too little information and are thus often incorrect. They were further told that although they would ultimately have to evaluate the effectiveness to each ratee's performance using seven-point scales, it was very important to postpone making judgments until they had adequately sampled the ratee's behavior. It was explained that focusing on only observing and counting relevant behaviors should help them not judge prematurely and hence rate more accurately. Once again, interviews with pilot test subjects indicated that this rationale for training seemed quite plausible, and actual experimental subjects raised no questions that suggested otherwise.

Trainees viewed the same two videotapes used in BARS training and practiced observing and rating the managers' performances. Subsequent to discussing what ratings had been given to each ratee, the trainer provided frequency feedback (i. e., how often each behavior

occured) to those who were rating with BOS and evaluative feedback (i. e., the correct scale value) to those who were rating with BARS. BOS training, like BARS training, was designed to last one and one-half hours.

Control Training. In order to keep the laboratory time constant for all groups, students who did not receive training participated in an one and one-half hour role play exercise (Maier, Solem, & Maier, 1957). Following role play, students were asked to observe and rate the videotaped performances. No discussion of the rating task was undertaken, other than to provide general instructions.

Dependent Variables

Items on the BOS that depicted ineffective manager performance were rescored prior to computing the accuracy measures so that large values on each dimension reflected better performance. For ratings on each format, then, higher values indicated more effective performance.

Although preference has been shown for a correlational index of accuracy as opposed to a difference score measure (Borman, 1979), the latter seemed more appropriate conceptually for the present research. This was because the focal question of interest was whether or not raters could be trained to match ratee performance to the most appropriate scale value on BARS and to report the frequency of critical behaviors on BOS. Such information could easily have been lost by assessing only the degree to which the true and observed scores covaried. Thus, two accuracy measures,

differential accuracy (DA) and distance from true scores (DIST), were computed; each is described in detail below.

Differential Accuracy. For each performance dimension, accuracy was computed using Cronbach's (1955) differential accuracy measure. This involved correlating each rater's ratings of the six videotaped target persons with the corresponding true scores. Fisher's r-to-z transformation was then applied to each DA correlation. For each subject, these analyses resulted in five z scores (one for each dimension) with higher scores indicating higher accuracy.

Distance from True Scores. A distance measure was computed that enabled consideration of "level, depression, and shape" (Nunnally, 1978, p. 442). Distance accuracy is the average absolute value of the deviation of the obtained ratings from the true scores. DIST was computed for subjects' ratings on each of the performance dimensions, with lower mean deviations indicating higher accuracy. The formulas used to calculate DIST for the BARS and BOS appear below. Slightly different calculations were necessary because distance was assessed at the dimension level for BARS and at the item level for BOS.

$$\text{BARS DIST} = \left(\sum_{r=1}^{R} D \right)/R$$

$$\text{BOS DIST} = \left[\sum_{r=1}^{R} \left(\sum_{i=1}^{I} D \right)/I \right]/R$$

where: DIST = accuracy score for each dimension across ratees.

R = number of ratees (6).

I = number of items per dimension

D = absolute difference of the observed score from the true score.

Interrater Agreement. In order to evaluate the effects of training on interrater agreement, intraclass correlations were computed for each dimension on the ratings made by subjects within each of the experimental groups. Tests of the differences between these intraclass correlations were done separately for those rating with BARS *versus those rating with* BOS.

Results

Relationships Between Rating Errors and Accuracy

Because of scaling differences between the BARS and BOS formats, comparisons across the scales were precluded with the two accuracy measures used here. Hence data were analyzed separately for the BARS format groups and for the BOS format groups. For each rating format, the means, standard deviations, and intercorrelations of subjects' accuracy scores, sex, age, and previous experience with performance appraisal are presented in Table 1. The two accuracy measures were highly intercorrelated for those rating with BARS as well as for those rating with BOS. Because identical training and dimension effects (presented below) were observed for the two accuracy measures, only the results for DIST (based on conceptual appropriateness) are presented.

Table 1. Means, SDs, and Intercorrelations of BARS and BOS Variables

Format	BARS							BOS	
	Mean	SD	(1)	(2)	(3)	(4)	(5)	Mean	SD
1. DA	1.01	.28		-.77	.07	.08	.04	.74	.17
2. DIST	1.14	.28	-.82		-.11	.05	.00	.34	.14
5. SEX	1.61	.50	.06	-.06		-.14	.08	1.50	.05
6. AGE	19.82	2.02	.05	-.09	.10		.04	19.01	1.27
7. EXP	1.49	.49	-.08	.06	.05	.00		1.45	.45

Note. Correlations greater than .18 are significant, $p < .05$. Correlations above the diagonal are for subjects who rated with BOS; those below the diagonal are for subjects who rated with BARS. DA = Correlation (transformed to z scores) between true and observed scores; DIST = Average difference between observed and true scores.

Training Effects for Behaviorally Anchored Rating Scales

The means and standard deviations for the DIST measures by training condition are presented in Table 2. A 3 x 5 (training x dimension) fixed-factor ANOVA was used to assess training and dimension effects. The dimension factor was a repeated measure. Results of that ANOVA revealed a significant main effect for training, $F(2, 69) = 33.02$, $p < .05$, $\omega^2 = .47$. Mean comparisons using Scheffe tests revealed that those who received congruent (i. e., BARS) training had significantly more accurate ratings than those who received BOS training or control training. Interestingly, ratings from the BOS training group were significantly more accurate than ratings from the control training group.

A significant main effect for dimension, $F(4, 276) = 13.74$, $p < .05$, $\omega^2 = .15$, and a significant training x dimension interaction, $F(8, 276) = 2.23$, $p < .05$, $\omega^2 = .03$, also resulted[1]. Analysis of the simple main effect (Winer, 1971) for training showed that on Structuring and Controlling the Interview and Resolving Conflict, BARS training yielded higher accuracy than BOS training, which, in turn, yielded higher accuracy than control training. For Motivating the Subordinate and Developing the Subordinate, there were no differences between the BOS training and control training groups, but ratings from both of these groups were significantly less accurate than ratings from the BARS training condition. Finally, no differences in accuracy resulted on Establishing and Maintaining Rapport.

Table 2. Means and SDs of DIST by Treatment

	BARS Format			BOS Format		
	BARS TRNG	BOS TRNG	NO TRNG	BARS TRNG	BOS TRNG	NO TRNG
DIM 1	.58	.95	1.57	.36	.18	.36
	(.20)	(.30)	(.46)	(.13)	(.12)	(.21)
DIM 2	1.23	1.23	.75	.45	.19	.42
	(.30)	(.35)	(.45)	(.12)	(.08)	(.12)
DIM 3	.87	1.22	1.20	.32	.23	.34
	(.29)	(.34)	(.37)	(.14)	(.05)	(.15)
DIM 4	.83	1.24	1.26	.53	.27	.56
	(.41)	(.34)	(.59)	(.17)	(.08)	(.16)
DIM 5	.91	1.25	1.22	.33	.14	.32
	(.25)	(.32)	(.40)	(.17)	(.06)	(.13)
Totals	.89	1.17	1.36	.40	.20	.41
	(.16)	(.16)	(.28)	(.12)	(.04)	(.11)

Note. The numbers in parentheses are standard deviations. Dimensions are: 1 = Structuring and Controlling the Interview; 2 = Establishing and Maintaining Rapport; 3 = Resolving Conflict; 4 = Motivating the Subordinate; 5 = Developing the Subordinate.

To evaluate training effects on interrater agreement, five (one per dimension) intraclass correlations were computed on the ratings from each experimental condition (see Table 3). Although a significant difference resulted for only one pair of correlations, a consistent pattern resulted across the groups. For all dimensions, those who received congruent (BARS) training had somewhat higher agreement (mean $\underline{r}$ across dimensions = .70) than those who received incongruent (BOS) training (mean $\underline{r}$ = .62), who, in turn, had higher agreement than the control training subjects (mean $\underline{r}$ = .50).

Training Effects for Behavioral Observation Scales

For those groups rating with BOS, the means and standard deviations for DIST by treatment are also shown in Table 2. The analysis aimed at evaluating training and dimension effects employed the accuracy measure in a 3 x 5 ANOVA, with training and dimension (repeated measures) as fixed-factors. The results of this ANOVA revealed a significant main effects for training, $\underline{F}(2, 69) = 34.07$, $\underline{p} < .05$, $\omega^2 = .49$. Scheffe tests showed that training congruent with the BOS format yielded significantly higher accuracy than did BARS training or control training. Further, not only were there no differences between the BARS and control training groups, but the means for these conditions were virtually identical.

A significant main effect for dimension, $\underline{F}$ (4, 276) = 33.03, $\underline{p} < .05$, $\omega^2 = .29$, and a significant training x dimension interaction, $\underline{F}(8, 276) = 3.39$, $\underline{p} < .05$, $\omega^2 = .04$, were also observed. Analysis of the simple main effect for training indicated that BOS training

Table 3. Intraclass Correlations by Treatment

	BARS Format			BOS Format		
	BARS TRNG	BOS TRNG	NO TRNG	BARS TRNG	BOS TRNG	NO TRNG
DIM 1	.82	.70	.47	.75	.85	.69
DIM 2	.69	.62	.52	.58	.69	.50
DIM 3	.61	.57	.40	.57	.68	.54
DIM 4	.65	.51	.55	.63	.72	.60
DIM 5	.73	.68	.61	.71	.87	.75
Mean *r*	.70	.62	.50	.64	.72	.62

Note. Dimensions are: 1 = Structuring and Controlling the Interview; 2 = Establishing and Maintaining Rapport; 3 = Resolving Conflict; 4 = Motivating the Subordinate; 5 = Developing the Subordinate. Significantly different correlations: for DIM 1 (BARS format), BARS TRNG > NO TRNG.

produced more accurate ratings than BARS training or control training on all performance dimensions. There were no differences between the latter two groups. The interaction was caused by relatively minor differences between the incongruent and control groups, which were inconsistent across performance dimensions.

Finally, intraclass correlations were computed to evaluate rater agreement within each of the BOS format groups (see Table 3). Although there were no significant differences between these correlations, a consistent pattern again resulted across the groups. Interrater agreement was higher for all dimensions within the BOS training group (mean $\underline{r}$ across dimensions = .76) than within both the BARS training group (mean $\underline{r}$ = .64) and the control group (mean $\underline{r}$ = .62). It is worthwhile to note that given the small samples (N = 24 per condition) used in these analyses, only quite large differences between the correlations would have resulted in significant differences.

Discussion

The results of this study suggest that rating accuracy can be increased by training individuals in a manner that is consistent with the cognitive demands of the particular rating format used. Specifically, when a BOS format was used to evaluate performance, only congruent (BOS) training was effective for increasing rater accuracy. Use of an incongruent (BARS) training strategy with BOS had no effect whatsoever on accuracy. For those using BARS, accuracy was also highest when training was congruent with the format. Interestingly, however, BOS training produced higher accuracy than no training when

BARS were used to evaluate performance. Although this result was not hypothesized, it does not seem unreasonable. Given the focus of BOS training, trainees most likely were able to accurately recall which behaviors they had observed each manager exhibit. Further, since the critical incidents used in BOS training did correspond to dimension anchors that appeared on the BARS, it seems reasonable to expect that raters may have been able to match their observations to a generally appropriate effectiveness level, resulting in fairly accurate ratings. Unlike BARS training, however, BOS training probably did not enable raters to make fine discriminations within general effectiveness levels.

The converse (*i. e.*, that BARS training would increase accuracy on a BOS) is less likely. Recall that the goal of BARS training was to teach raters a common set of evaluative standards. This was accomplished by providing prototypical examples of what constituted effective, average, and ineffective manager performance on each rating dimension. Because of this training, raters probably did observe the videotapes with an evaluative orientation. When some trainees were subsequently required to report how frequently specific behaviors occurred using BOS, their recall of these behaviors was most likely structured according to their general evaluative judgments (Estes, 1976; Hamilton, Katz, & Leirer, 1980; Murphy et al., 1982). There is little, if any, reason to believe that these judgments could have been translated to accurate frequency counts. After all, once a ratee has been categorized, for example, as an effective performer, the features of the prototype to which s/he was matched come to characterize the

individual, making recall of very specific information more difficult to achieve (Cantor & Mischel, 1977, 1979; Wyer & Srull, 1980).

Training that is designed for a given rating format may facilitate accuracy through multiple mechanisms. First, by focusing on components of the rating process that are most salient in using a particular type of scale, relevant rater skills are developed and/or enhanced. Also, research has shown that the goals of the rating task influence information processing in terms of what data are sought, how they are stored (Cohen & Ebbesen, 1979), and what can be recalled (Hamilton, et al., 1980; Lingle, Geva, Ostrom, & Baumgardner, 1979). Congruent training seems to make clear the rating "goals" (observation versus evaluation, in this case), and may also have the effect of motivating raters to use the scales in common and appropriate ways. The finding that interrater agreement was consistently higher for each congruent training condition seems to support the notion that raters shared a common orientation to their particular rating task.

Questions remain, however, concerning whether or not the present training programs will facilitate accuracy in ratings made over relatively long time periods, for example, over six months or a year. Regardless of whether raters are using BARS, BOS, or some other format, it is unlikely that they will be able to remember large amounts of detailed information over time (Heneman & Wexley, 1982; Murphy, et al., 1982; Wyer & Srull, 1980). Nevertheless, the training strategies employed here seem potentially useful. Training with the dimensional system of the rating scales should promote a more complete

understanding of the performance domain and should also motivate raters to consider ratee performance in job relevant ways. Given that job-relevant categories are the source of valid variance in performance appraisals (Ilgen & Feldman, 1983), the types of training suggested here are, at a minimum, a first step towards accuracy.

Accuracy Training for Other Rating Tasks

Because the long term effects of the present training programs have yet to be evaluated in ongoing performance appraisal situations, the data reported here may be most relevant to short term observational tasks. Consider, for example, an assessment center situation in which ratings of assessee performance are typically made subsequent to each exercise and thus, recall demand are minimized. Assessment centers are also similar to the present laboratory conditions in the sense that they are void of many extraneous factors that most likely influence ongoing performance appraisals (e. g., purpose of the appraisal, rater/ratee interpersonal relationship, etc.). The training effects obtained in this study might thus be indicative of what could be expected if similar strategies were used to train assessors. Also, Sackett and Dreher (1982) have recently brought into question the stability of assessment center dimensions across exercises. It seems possible that dimension instability may be due, at least in part, to the large inferences that are often required of assessors as they move from observing behavior in exercises to making overall ratings on complex dimensions such as leadership and analytical skills (Sackett & Dreher, 1984). A possible solution for

decreasing the amount of inference necessary may lie in developing more behaviorally-oriented rating scales, which tap similar (or even different) dimensions to those already used in assessment centers. Such rating instruments coupled with the types of training suggested here may well yield more stable dimension ratings across exercises.

Another rating situation in which accuracy training of the type used here might prove especially useful is the employment interview. Again, this rating task does not require recall of information over time, and it is also not plagued by many of the extraneous factors that influence appraisals. Previous research has shown the importance of providing interviewers with job-relevant, dimensional rating scales for evaluating applicants (Osburn, Timmreck, & Bigby, 1981). As the present study would suggest, however, merely providing raters with an appropriate format does not ensure its proper use. Accuracy training, similar to that employed here, may be useful in this regard.

Limitations and Future Research Directions

Although the hypotheses set forth in this study were supported, there are limitations that should be recognized in drawing conclusions based on these data. First, undergraduate students and not managers were used as subjects and consequently, the results can only tentatively be generalized to a true manager population. Observations were also made from videotaped rather than live performances. Thus, this study could be replicated and extended by using more experienced raters, live performances, and perhaps other training procedures. Also important is the need to evaluate the stability of the accuracy

training effects over time. Finally, it is worthwhile to note that the present training programs were both developed for behaviorally-based rating instruments. Whether or not training could be designed to increase accuracy with other types of formats (e. g., trait) is an interesting question for future research.

Conclusions

On a practical level, the results of this study suggest that there is no "one best way" to train raters to make accurate performance ratings. Rather, training should be dictated by the particular demands placed on raters by the format being used. The question of which format is superior may best be answered by considering relevant aspects of the performance domain to be evaluated. For example, BARS might be more appropriate for performance dimensions that are inherently evaluative or difficult to define in terms of concrete, observable behaviors. For other rating content, BOS items may be preferable. Irrespective of which format is used, however, the data reported here indicate that rating accuracy can be facilitated when training is developed in accordance with the rating scales.

Reference Notes

Frankman, R. Personal communication (March 1983).

References

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors of student ratings of instructors. Journal of Applied Psychology, 63, 125-131.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 61, 564-570.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bernardin, H. J., & Kane, J. (1980). A second look at behavioral observation scales. Personnel Psychology, 33, 809-814.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 233-252.

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C. (1983). Implications of implicit personality theory and personal constructs for the rating of work performance in organizations. In F. J. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Erlbaum.

Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. Journal of Personality and Social Psychology, 35, 38-48.

Cantor, N., & Mischel, W. (1972). Prototypes in person perception. In L. Berkowitz (Ed.), Advances in experimental social psychology, Vol. 12. New York: Academic Press.

Cohen, C. E., & Ebbesen, E. B. (1979). Observational goals and schema activation: A theoretical framework of behavior perception. Journal of Experimental Social Psychology, 15, 305-329.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 52, 177-193.

Estes, W. (1976). The cognitive side of probability learning. Psychological Review, 83, 37-64.

Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Cognitive representation of personality impressions: Organizational processes in first impression formation. Journal of Personality and Social Psychology, 39, 1050-1063.

Heneman, R. L., & Wexley, K. N. (1982). The effects of time delay in rating and amount of information observed in performance rating accuracy. Academy of Management Journal, 26, 677-687.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process approach. In B. M. Staw & L. L. Cummings (Eds.), Research in organizational behavior, Vol. 5. Greenwich, Conn.: JAI Press Inc.

Landy, F. J., & Farr, J. (1970). Performance rating. Psychological Bulletin, 87, 72-107.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30, 255-268.

Latham, G. P., & Wexley, K. N. (1980). Increasing productivity through performance appraisal. Reading, Mass.: Addison-Wesley Publishing Co.

Latham, G. P., Wexley, K. N., & Purcell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lingle, J. H., Geva, N., Ostrom, T. M., Leippe, M. R., & Baumgardner, M. H. (1979). Thematic effects of person judgments on impression organization. Journal of Personality and Social Psychology, 37, 674-687.

McIntyre, R., Smith, D., & Hassett, C. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Maier, N. R. F., Solem, L., & Maier, E. (1957). Supervisory and executive development. New York: Wiley & Sons.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67, 262-267.

Nunnally, J. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Osburn, H. G., Timmreck, C, & Bigby, D. (1981). Effect of dimensional relevance on accuracy of simulated hiring decisions by employment interviewers. Journal of Applied Psychology, 66, 159-165.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, in press.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67, 401-410.

Sackett, P. R., & Dreher, G. F. (1984). Situational specificity and assessment center validation strategies: A rejoinder to Nedig and Nedig. Journal of Applied Psychology, 69, 401-410.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. (1975). Behavioral anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Winer, B. J. (1971). Statistical principles in experimental design 2nd ed. New York: McGraw-Hill.

Wyer, R. S., Jr., & Srull, T. K. (1980). Category accessibility: Some theoretical and empirical issues concerning the processing of social stimulus information. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), Social cognition: The Ontario symposium on personality and social psychology. Hillsdale, N. J.: Erlbaum.

Footnotes

[1] The omega squares were computed separately for the between and and within subjects effects and thus are not directly comparable (Frankman, Note 1).

LIST 1 MANDATORY*

Defense Technical Information Center (12)
ATTN: DTIC DDA-2
Selection & Preliminary Cataloging Section
Cameron Station
Alexandria, VA 22314

Library of Congress
Science and Technology Division
Washington, D.C. 20540

Office of Naval Research (3)
Code 4420E
800 N. Quincy Street
Arlington, VA 22217

Naval Research Laboratory (6)
Code 2627
Washington, D.C. 20375

Office of Naval Research
Director, Technology Programs
Code 200
800 N. Quincy Street
Arlington, VA 22217

LIST 2 ONR FIELD

Psychologist
Office of Naval Research
Detachment, Pasadena
1030 East Green Street
Pasadena, CA 91106

LIST 3 OPNAV

Deputy Chief of Naval Operations
(Manpower, Personnel, & Training)
Head, Research, Development, and
Studies Branch (Op-115)
1812 Arlington Annex
Washington, D.C. 20350

Director
Civilian Personnel Division (OP-14)
Department of the Navy
1803 Arlington Annex
Washington, D.C. 20350

Deputy Chief of Naval Operations
(Manpower, Personnel, & Training)
Director, Human Resource Management
Plans & Policy Branch (OP-150)
Department of Navy
Washington, D.C. 20350

LIST 4 NAVMAT & NPRDC

Program Administrator for Manpower,
Personnel, and Training
MAT-0722
800 N. Quincy Street
Arlington, VA 22217

Naval Material Command
Management Training Center
NAVMAT 09M32
Jefferson Plaza, Bldg #2, Rm 150
1421 Jefferson Davis Highway
Arlington, VA 20360

Naval Material Command
Director, Productivity Management Office
MAT-00K
Crystal Plaza #5
Room 632
Washington, D.C. 20360

Naval Personnel R&D Center (4)
Technical Director
Director, Manpower & Personnel
Laboratory, Code 06
Director, System Laboratory, Code 07
Director, Future Technology, Code 41
San Diego, CA 92152

*Number in parentheses is the number of copies to be sent.

Navy Personnel R&D Center
Washington Liaison Office
Ballston Tower #3, Room 93
Arlington, VA 22217

LIST 5 BUMED

NONE

LIST 6
NAVAL ACADEMY AND NAVAL POSTGRADUATE SCHOOL

Naval Postgraduate School (3)
ATTN: Chairman, Dept of
Administrative Science
Department of Administrative Sciences
Monterey, CA 93940

U.S. Naval Academy
ATTN: Chairman, Department of
Leadership and Law
Stop 7-B
Annapolis, MD 21402

LIST 7 HRM

Officer in Charge
Human Resource Management Division
Naval Air Station
Mayport, FL 32228

Human Resource Management School
Naval Air Station Memphis (96)
Millington, TN 38054

Commanding Officer
Human Resource Management School
Naval Air Station Memphis
Millington, TN 38054

LIST 8 NAVY MISCELLANEOUS

Naval Military Personnel Command (2)
HRM Department (NMPC-6)
Washington, D.C. 20350

LIST 9 USMC

Headquarters, U.S. Marine Corps
ATTN: Scientific Adviser,
Code RD-1
Washington, D.C. 20380

LIST 10 OTHER FEDERAL GOVERNMENT

Dr. Brian Usilaner
GAO
Washington, D.C. 20548

Social and Developmental Psychology
Program
National Science Foundation
Washington, D.C. 20550

Office of Personnel Management
Office of Planning and Evaluation
Research Management Division
1900 E. Street, N.W.
Washington, D.C. 20415

LIST 11 ARMY

Technical Director (3)
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Head, Department of Behavior
Science and Leadership
U.S. Military Academy, New York 10996

LIST 12 AIR FORCE

Air University Library
LSE 76-443
Maxwell AFB, AL 36112

Head, Department of Behavioral
Science and Leadership
U.S. Air Force Academy, CO 80840

LIST 13 MISCELLANEOUS

Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, VA 22207

LIST 14 CURRENT CONTRACTORS

Dr. Janet L. Barnes-Farrell
Department of Psychology
University of Hawaii
2430 Campus Road
Honolulu, HI 96822

Jeanne M. Brett
Northwestern University
Graduate School of Management
2001 Sheridan Road
Evanston, IL 60201

Dr. Terry Connolly
Georgia Institute of Technology
School of Industrial & Systems
Engineering
Atlanta, GA 30332

Dr. Richard Daft
Texas A&M University
Department of Management
College Station, TX 77843

Dr. Randy Dunham
University of Wisconsin
Graduate School of Business
Madison, WI 53706

Dr. Lawrence R. James
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332

Dr. J. Richard Hackman
School of Organization & Management
Box 1A, Yale University
New Haven, CT 06520

Dr. Frank J. Landy
The Pennsylvania State University
Department of Psychology
417 Bruce V. Moore Building
University Park, PA 16802

Dr. Bibb Latane
The University of North Carolina
at Chapel Hill
Manning Hall 026A
Chapel Hill, NC 27514

Dr. Edward E. Lawler
University of Southern California
Graduate School of Business Administration
Los Angeles, CA 90007

Dr. William H. Mobley
College of Business Administration
Texas A&M University
College Station, TX 77843

Dr. Thomas M. Ostrom
The Ohio State University
Department of Psychology
116E Stadium
404C West 17th Avenue
Columbus, OH 43210

Dr. Robert Rice
State University of New York at Buffalo
Department of Psychology
Buffalo, NY 14226

Dr. Benjamin Schneider
Department of Psychology
University of Maryland
College Park, MD 20742

Dr. H. Wallace Sinaiko
Program Director, Manpower Research
and Advisory Services
Smithsonian Institution
801 N. Pitt Street, Suite 120
Alexandria, VA 22314

Dr. Richard M. Steers
Graduate School of Management
University of Oregon
Eugene, OR 97403

Dr. Harry C. Triandis
Department of Psychology
University of Illinois
Champaign, IL 61820

Dr. Anne S. Tsui
Duke University
The Fuqua School of Business
Durham, NC 27706

Andrew H. Van de Ven
University of Minnesota
Office of Research Administration
1919 University Avenue
St. Paul, MN 55104


DAT
FILM
7