Scaling Behavioral Anchors

Frank J. Landy and Janet L. Barnes
The Pennsylvania State University

Although behaviorally anchored rating scales (BARS) have both intuitive and empirical appeal, they have not always yielded superior results in contrast with graphic rating scales. The present study examined the issue of how behavioral descriptions are anchored. Subjects scaled anchors describing teaching performance in a college classroom using either a graphic rating procedure or a pair-comparison procedure. The two different methods resulted in scale anchors with different properties, particularly with respect to item dispersions. It was proposed that the choice of an anchoring procedure depends on the nature of the actual rating process.

Since behaviorally anchored rating scales (BARS) were introduced by Smith and Kendall (1963), they have commanded a good deal of the attention of researchers concerned with the issue of the reliability and validity of performance ratings. Unfortunately, the results of this research have been equivocal. Recent reviews (Bernardin, 1977; Borman & Vallon, 1974; DeCotiis, 1977; Schwab, Heneman, & Decotiis, 1975) have indicated that the procedures suggested by Smith and Kendall (1963) may not be sufficient to insure ratings of higher quality than those which might have been obtained by the use of traditional graphic rating scales.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 2 Spring 1979 pp. 193-200

Unfortunately, many of the tests of differences between graphic scales and BARS have been confounded. On the one hand, when BARS are compared to poorly developed graphic scales, they seem to be superior (Campbell, Dunnette, Arvey, & Hellervik, 1973). On the other hand, when BARS are compared with well-developed graphic scales (Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976), the differences between the two different types of scales are less pronounced.

There have been a number of studies which have demonstrated that the logic of developing behaviorally anchored scales is sound:

1. There is some advantage to having the ultimate users of scales involved in their development (Friedman & Cornelius, 1976).
2. Behavioral scale anchors are more informative than simply numbers or adjectives (Barrett, Taylor, Parker, & Martens, 1958; Bendig, 1952; Peters & McCormick, 1966).
3. Behaviorally anchored scale construction yields anchors which have similar meanings to all raters (Campbell, Dunnette, Arvey, & Hellervik, 1973; Harari & Zedeck, 1973; Landy & Guion, 1970; Smith & Kendall, 1963).
4. Behaviorally anchored performance dimensions can be operationally and conceptually distinguished from one another (Campbell et al., 1973; Smith & Kendall, 1963).

The entire developmental process of BARS construction is predicated on the proposition that unambiguous performance dimensions, defined and anchored unequivocally in the language of the rater, will provide a reliable and accurate vehicle for rater decision making concerning the ratee's levels of performance.

Although the scale construction process is sound in its conception, there are a few areas in which improvement might be made. Several studies have demonstrated that the resistance of rating scales to typical errors, such as leniency and halo, is partially a function of the level of psychometric rigor which characterized the initial scale development. In their original research, Smith and Kendall (1963) demonstrated the value of item analysis procedures for the selection of potential scale anchors. Bernardin (Bernardin, 1977; Bernardin et al., 1976) has demonstrated that the BARS format compares favorably with other methods when rigorous item analysis procedures are used in item selection. In the BARS methodology, there are several steps which comprise the process of "item selection." The first step is to select items which help distinguish good performers from poor performers; this process is implied in the term "critical incidents" so often used to characterize BARS development. The second step is to find items which form a dimension with as little overlap as possible on another dimension. The third step is to scale these items on the relevant dimension. The scaling involves assigning a particular value to an item which represents the relative desirability of that particular item. It is this scaling process which is examined in the present research.

In the most common form of BARS development, judges are requested to assign scale values ranging from 1 to 9 (or 1 to 7) to behavioral examples. These numbers are intended to represent the desirability of the examples. Through this process, examples are arranged on an "order of merit" continuum. This type of anchoring might be contrasted with more sophisticated procedures. Several studies have demonstrated that anchoring procedures may affect the resulting means and standard deviations of behavioral examples (Landy & Guion, 1970; Rotter & Tinkleman, 1970; Wells & Smith, 1960). In addition, there is a rich body of theory and data which suggests that various scaling procedures produce differential results (Edwards, 1957; Guilford, 1954; Torgerson, 1958). Yet the issue of stimulus scaling has received little attention in BARS research.

Judges in the BARS development procedure are typically requested to make absolute judgments about the desirability of the potential anchors. Nunnally (1967) and others have suggested that comparative judgments are superior to absolute judgments. Thurstone's Law of Comparative Judgment (Guilford, 1954) describes the operations which might be used to anchor items comparatively rather than absolutely. One of the most common techniques for comparative scaling is the pair-comparison technique. The present study contrasts scale values assigned to behavioral anchors using the typical BARS absolute judgment technique with scale values derived from a pair-comparison presentation of behavioral anchors.

Method

Subjects

Subjects were 299 students recruited from introductory psychology classes. Each subject was offered extra credit for participation in the study. In addition to sex, information about major, term standing, and previous experience with teacher evaluations was requested of each subject.

Materials

Two sets of statements about teacher behaviors were chosen from among a number of behavioral expectation scales previously developed by Harari and Zedeck (1973). Dimension I, *Ability to Motivate,* was composed of 10 statements

about the instructor's ability to generate student interest in subject matter. Dimension II, *Delivery,* consisted of 9 statements about the instructor's ability and way of conveying the material. The items appear in Table 1.

Procedure

Students were randomly assigned to one of two conditions: pair comparison (Treatment A) or graphic rating (Treatment B). The 148 students assigned to Treatment A were presented with two standard pair comparison rating tasks. The 10 statements comprising Dimension I were paired in all possible combinations; students were instructed to examine each pair of statements carefully and, for each pair, to decide which of the two statements represented "more" of the dimension being rated. The same procedure was followed in making judgments about the 9 statements constituting Dimension II. The order of statements was counterbalanced across presentations. The 151 students assigned to Treatment B were asked to provide judgments about the statements constituting Dimension I and Dimension II using graphic rating scales. Statements were presented individually; students were requested to decide "how much" of the dimension described was represented by each statement and to assign a scale value from 1 (minimum amount) to 7 (maximum amount) to each statement. The data from these judgments were used to derive mean scale values for each of the behavioral statements under Treatment A and Treatment B.

Data Analysis

Scale values and standard deviations were computed for Treatment A using the classic Thurstone solution. Since the standard deviations of the stimulus elements were heterogeneous (Dimension I: $x^2 = 289.97$, $p < .01$; Dimension II: $x^2 = 136.03$, $p < .01$), a Case III solution was applied (Guilford, 1954). Mean scale values and their standard deviations were also computed for Treatment B. In order to make direct comparisons of scale values and variances for each item possible, the mean and the standard deviation of each *set* of scale values were calculated (Dimension I: $n = 10$; Dimension II: $n = 9$), and these were used to standardize mean scale values and standard deviations for treatments.

Differences in mean scale values for each item were tested using a Behrens-Fisher t for independent samples, with Welch degrees of freedom. The familywise risk of Type I error was controlled through the use of the Bonferroni t procedure. Differences in variances for each item were tested using the F-test for variances, with the familywise risk of Type I error controlled via the Bonferroni procedure.

Results

An examinination of the covariates led to the conclusion that there were no differences due to sex, major, term standing, or experience with teacher evaluations; therefore, data were collapsed across these categories, and all analyses were conducted on the total sample. Since means and standard deviations were available from the original Harari and Zedeck (1973) research, comparisons were made among the three sets of scale values and standard deviations. The means and standard deviations appear in Table 2.

The results of the comparisons appear in Table 3. The results support the hypothesis that scaling procedures affect final item scale values and variances. In addition, the comparison of the scale values and variances from the Harari and Zedeck (1973) study with the corresponding values from Treatment B of the present study indicates little difference in item variance, but several significant differences in item scale values.

Discussion

The results of the present study are mixed with respect to the traditional procedures for de-

Table 1

Behavioral Anchor Items for Ability to Motivate and Delivery Dimensions

Ability to Motivate

A. This professor could be expected to be so inspiring that the student is often ahead in his reading assignments.
B. After completing an introductory social psychology course with this professor, most students could be expected to enroll in other classes that deal with the field of social psychology.
C. In this developmental psychology class, if a student hesitatingly describes a little experiment with school children that he is thinking about, this professor could be expected to reply: "Great! It sounds good. Your plan has some flaws, but every psychologists' plan has some flaws at first. We can work it out, and I'm sure you'll enjoy doing it!"
D. In an introductory psychology class, this professor could often be expected to pose questions and issues to students that are later discussed in section meetings or with classmates and friends outside the class.
E. This professor's students could be expected to have no qualms about studying the material he assigns.
F. The students in this professor's class could be expected to do the required work.
G. The students in this professor's class could be expected to do the required work and no more.
H. This professor of a psychological statistics class could be expected to try to push students into being interested by almost pleading with them.
I. Attendance in this professor's class could be expected to be less than 50% each meeting.
J. After completing an introductory psychology course with this professor, most students could be expected to be so disillusioned with psychology that they have little desire to enroll in other psychology courses.

Delivery

A. This professor could be expected to have a clear, excellent voice and can be heard anywhere in the auditorium. He could be expected to speak with inflection and to convey each mood of the material.
B. This professor's use of visual aids could be expected to entertain and inform the students.
C. This professor, when contrasting operant and classical conditioning, could be expected to make good use of the blackboard.
D. This professor's voice could be expected to be clear and distinct but sometimes he could be expected to speak too fast for the student to get the material into his notes.
E. In this introductory psychology class, students could be expected to have no difficulty understanding this professor's lecture on conditioned-response sets, but they could often be expected to be bewildered when he discusses theory in general.
F. When lecturing, this professor could be expected to pace across the platform back and forth and make the students nervous.
G. On occasion, this professor could be expected to mumble to himself in the middle of a lecture.
H. In order to study for an exam of this professor's, students could be expected to go the the TAs because they can't understand the explanations of the professor.
I. This professor could be expected to read from his notes and to speak in a low monotone. It is almost impossible not to become drowsy in class.

Table 2

Item Mean Scale Values and Standard Deviations[a]

Item	Pair Comparison (A) $\bar{X}$	Pair Comparison (A) SD	Graphic (B) $\bar{X}$	Graphic (B) SD	Harari and Zedeck (C) $\bar{X}$	Harari and Zedeck (C) SD
Ability to Motivate						
A.	1.86	1.66	1.52	.54	1.53	.58
B.	1.05	.66	1.21	.55	1.29	.62
C.	.90	.74	1.06	.67	1.09	.61
D.	.26	.47	.55	.59	.28	.60
E.	- .10	.37	- .16	.61	- .13	.78
F.	- .26	.50	- .39	.59	- .39	.70
G.	- .88	.77	- .84	.64	- .54	.61
H.	- .89	1.00	- .86	.63	- .94	.65
I.	-1.00	1.07	-1.03	.64	-1.05	.54
J.	- .95	1.02	-1.07	.59	-1.13	.60
Delivery						
A.	1.49	.13	1.06	.45	1.26	.54
B.	.93	1.22	.95	.53	1.00	.60
C.	.49	.57	.64	.51	.76	.59
D.	.40	.54	.55	.58	.58	.60
E.	.24	.73	.01	.54	.02	.55
F.	- .10	.60	- .07	.60	- .06	.57
G.	- .78	.61	- .86	.52	- .76	.59
H.	- .93	.76	- .94	.56	-1.24	.62
I.	-1.74	1.34	-1.33	.59	-1.55	.48

[a]Based on a scale where $\bar{X}$ = 0.00, SD = 1.00

Table 3
Comparison of Item Means and Variances Across Treatments

Item	Means			Variances		
	A - B	A - C	B - C	A - B	A - C	B - C
Ability to Motivate						
A.				*	*	
B.		*				
C.						
D.	*		*		*	
E.				*	*	*
F.					*	
G.		*	*			
H.				*	*	
I.				*	*	
J.				*	*	
Delivery						
A.	*	*	*	*	*	
B.				*	*	
C.		*				
D.						
E.	*	*		*	*	
F.						
G.						
H.		*	*	*		
I.	*		*	*	*	

veloping behaviorally anchored scales. As was mentioned in the results section, a test of the variance for the pair comparison scaling technique suggested that these variances were heterogeneous, thus requiring a Case III solution. An examination of Table 2 shows that this was not the case for either of the graphic scaling techniques; a test of these dispersions showed that they were homogeneous. With the present data, it is impossible to determine which technique provides "correct" values, but the fact remains that the sets of variances were clearly different.

Additionally, the item variances can be tested for *each* item across methods. These results are presented in Table 3 and suggest once again that the two techniques (pair comparison and graphic) result in anchors with different dispersions. Almost all of the significant differences between item variances involved contrasts between the graphic and pair-comparison methods; in only one instance was there a difference between the Harari and Zedeck variances and the present graphic variances. This strongly suggests that it is the method of scaling which is introducing the differences, rather than differences in subject populations.

A comparison of mean values presents a somewhat different picture. There are relatively few serious discrepancies between the graphic mean scale values and the pair-comparison values. It might reasonably be concluded that the graphic scaling technique is remarkably robust, yielding values comparable to those obtained from a rather sophisticated scaling method.

There are other, more subtle, differences which can be seen from a closer inspection of Table 2. In the pair-comparison procedure, the end anchors had substantially higher variances than the anchors in the center of the scale. This is an uncommon finding with the pair-comparison technique. Once again, it is impossible to determine if this is due to properties of the scaling procedures or to properties of the anchors themselves. Nevertheless, it should be examined in future research. It is possible that due to its absolute judgment nature, the graphic scaling procedure produces dispersions which are inaccurate estimates of population values. This would mean that many items which are included as final anchors might have been excluded had a different scaling procedure been used.

It also appears that one effect of the pair-comparison procedure was to "stretch out" the positive ends of the two rating scales. This can be seen from the fact that in both cases the pair comparison mean value for anchor A was more extreme than the comparable graphic values. This finding should not be overinterpreted, since the larger standard deviation for the pair-comparison values made the test for the significance of the differences nonsignificant. Nevertheless, an examination of the scale values suggests that certain items may be displaced by one or the other scaling technique.

In terms of the "reliability" of the two solutions, it might be argued that the pair-comparison results should be more stable because there are substantially more data points per anchor for that technique than for the graphic method. This would lead to the conclusion that the pair-comparison dispersions are better estimates of the population counterparts than the graphic estimates. Perhaps a more salient issue might be the manner by which one person judges the level of performance of a second person. If the judge compares anchors and attempts to place a person on a performance continuum by means of that comparative operation, then scale values and dispersions generated through pair-comparison procedures might be more representative. On the other hand, if a judge attempts to "match" an individual with an anchor on an anchor-by-anchor basis, then perhaps the graphic mean and dispersions are more reasonable. This implies that much more must be known about the way in which a person *uses* a rating scale to make a judgment before an anchor scaling method can be chosen.

Another issue of some interest is that of the items which were eliminated by Harari and Zedeck on the basis of high standard deviations. Since they were not available for the present study, it is impossible to determine whether or not they would have been dropped on the basis of pair-comparison scaling. It is likely that items in the center of the scale which were dropped by Harari and Zedeck might not have been eliminated if the pair-comparison technique had been used. That would not have been the case for items toward the ends of the scales. The present data suggest that any extreme item which was dropped by Harari and Zedeck would also have been eliminated by pair-comparison scaling. This is a matter of some importance, since one of the major flaws in BARS construction has been an inability to identify items in the center of scales with sufficiently low standard deviations to warrant inclusion in the final scale. It may be that pair-comparison scaling eliminates that problem in the construction of behaviorally anchored scales.

In summary, it appears that the two different scaling techniques produce substantially different results. It is impossible to identify which scaling procedure is more "correct" or "accurate" at this point. This will depend on a clear understanding of the process by which raters utilize scales for making decisions. Nevertheless, the present results suggest that much more research needs to be done on the procedure which is used to assign scale values to behavioral anchors. Through this research it may be possible to produce rating instruments which are more compatible with rater behaviors than current developmental techniques allow.

References

Barrett, R., Taylor, E., Parker, J., & Martens, W. Rating scale content: I. Scale information and supervisory ratings. *Personnel Psychology*, 1958, *11*, 333–346.

Bendig, A. W. A statistical report on a revision of the Miami instructor rating sheet. *Journal of Educational Psychology*, 1952, *43*, 423–429.

Bernardin, H. J. Behavioral expectation scales versus summated scales: A fairer comparison. *Journal of Applied Psychology*, 1977, *62*, 422–426.

Bernardin, H. J., Alvares, K., & Cranny, C. A recomparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*, 1976, *6*, 564–570.

Borman, W., & Vallon, W. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology*, 1974, *59*, 197–201.

Campbell, J., Dunnette, M., Arvey, R., & Hellervik, L. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, *57*, 15–22.

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

Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.

Friedman, B. A., & Cornelius, E. T. Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. *Journal of Applied Psychology*, 1976, *61*, 210–216.

Guilford, J. *Psychometric methods*. New York: McGraw-Hill, 1954.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. *Journal of Applied Psychology*, 1973, *58*, 261–265.

Landy, F., & Guion, R. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, *5*, 93–103.

Nunnally, J. *Psychometric theory*. New York: McGraw-Hill, 1967.

Peters, D., & McCormick, E. Comparative reliability of numerically anchored versus job–task anchored rating scales. *Journal of Applied Psychology*, 1966, *50*, 92–96.

Rotter, G., & Tinkelman, V. Anchor effects in the development of behavior rating scales. *Educational and Psychological Measurement*, 1970, *30*, 311–318.

Schwab, D., Heneman, H., & DeCotiis, T. Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, 1975, *28*, 549–562.

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

Torgerson, W. *Theory and methods of scaling*. New York: Wiley, 1958.

Acknowledgments

We gratefully acknowledge the help of Kit Ford, Cheryl Mannino, and Jeff Speidel in data collection and preparation, as well as the suggestions of Ian Spence and Hoben Thomas regarding data analysis. Sheldon Zedeck graciously supplied the scales used in the Harari and Zedeck (1973) study, as well as the means and standard deviations for the anchors.

Author's Address

Send requests for reprints or further information to Frank J. Landy, Department of Psychology, The Pennsylvania State University, University Park, PA 16802.