A Multitrait-Multirater Analysis of a Behaviorally-Anchored Rating Scale for Sales Personnel

John M. Ivancevich
University of Houston

Behaviorally-anchored rating scales *(BARS)* have grown in popularity among researchers and practitioners. There have been reports and studies of *BARS* being used to evaluate the performance of nurses, engineers, grocery clerks, managers, and teachers. An important issue in using *BARS* concerns the validity of the instrument. This study of the development and validity evaluation by the multitrait-multirater approach of a *BARS* for sales personnel suggests (1) moderate convergent validity and (2) little or no discriminant validity.

Recently there appears to be an increased interest in studying and applying behavioral-specific measures for evaluating the performance of employees (Bernardin & Walter, 1977; Campbell, Dunnette, Lawler, & Weick, 1970; Fogli, Hulin, & Blood, 1971; Williams & Seiler, 1973). These scales are typically constructed using an iterative procedure described by Smith and Kendall (1963). The purpose of this paper is to evaluate the convergent and discriminant validity of a behaviorally-anchored rating scale developed for sales personnel. The multitrait-multirater approach is used to provide a more sophisticated understanding of behavioral-specific measures than is possible where it is not employed (Campbell & Fiske, 1959).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 4 Fall 1977 pp. 523–531

The procedure used to develop behaviorally-anchored rating scales (*BARS*) is a variant of critical incident methodology that requires the appropriate organizational personnel to consider in depth the components of performance for the job being studied. These personnel also define anchors for the performance criteria in specific and relevant behavioral terminology. This process of developing *BARS* is assumed to have a number of important advantages over traditional performance appraisal techniques such as graphic rating scales. Some of the assumed advantages are: (1) individuals with work experience similar to those who eventually use the scales participate in the actual construction of the scales; (2) critical behavioral incidents are used as anchor points on each scale; (3) the terminology used for describing the job is retained in the anchors; and (4) relatively independent scales with high scale reliabilities are obtained (Smith & Kendall, 1963).

These advantages are certainly appealing to individuals who, as part of their responsibilities, must rate subordinates or peers. There have been reports of using *BARS* for evaluating nurses (Goodale & Burke, 1975; Smith & Kendall, 1963; Zedeck & Baker, 1972), managers (Campbell, Dunnette, Arvey, & Hellervik, 1973), engineers (Williams & Seiler, 1973), grocery clerks (Fogli, et al., 1971), teachers (Harari & Zedeck, 1973), and programmer and system analysts (Arvey & Neel, 1974).

Zedeck and Baker (1972), using the Smith and Kendall (1963) scale, studied head nurses and supervisory nurse appraisals of registered nurse performance. They used a multitrait-multirater analysis and found moderate convergent validity and no discriminant validity for the performance dimensions.

Dickinson and Tice (1973) investigated the extent to which rating-scale dimensions obtained by the re-translation phase of the Smith and Kendall (1963) procedure are used differentially by raters. The *BARS* was developed by professional firefighters and their immediate supervisors at four municipal fire departments. A multitrait-multimethod analysis of the ratings indicated that the re-translation procedure resulted in dimensions with little discriminant validity and only a moderate degree of convergent validity.

The intent of the present study was to develop and evaluate various psychometric properties of *BARS* for field sales personnel. Since performance appraisal of sales personnel is so important, it seems necessary to utilize only valid and reliable methods. A multitrait-multirater analysis (Campbell & Fiske, 1959; Lawler, 1967) was used to assess the validity and reliability of the occupation- and organization-specific *BARS*.

Method

Participants

The study was conducted in a large national organization. The executives of the company were dissatisfied with the traditional appraisal program, which was discarded six months prior to the start of the study. The company's total sales unit contained 310 sales people who reported to 44 district sales managers. The district managers reported to regional managers. A six-phase development and evaluation program was employed since the company's management wanted to include as many of the sales personnel and managers as possible in the actual development of the final *BARS*.

Development of the BARS

Phase I. Six district managers and six sales people were randomly selected from the geographical districts. This group met for two days and discussed the performance rating process and the critical incident procedure. Next, the group was given the task of identifying independent job performance dimensions for the sales persons' jobs and to define them by writing three general descriptor (critical incident) statements.

Phase II. A new group of six district managers and six sales personnel met for two days. The first task of the group was to review critically, but constructively, the output generated in Phase I. The second task of the group was to write specific descriptors (critical incidents) for each performance dimension which they wanted to retain.

Phase III. A new group of 30 sales people, six district managers, and one regional manager met for one day and discussed the results of the Phase I and II portions of the program. After being given a list of the specific descriptor statements, they assigned each descriptor (critical incident) to a performance dimension. This part of the development of the *BARS* is designed to assess the independence and clarity of the dimensions and to identify the ambiguous descriptors which could cause problems when used by raters. A specific descriptor was retained if 25 (or approximately 70 percent) of the Phase III participants assigned it to the same dimension.

Phase IV. Specific descriptors retained through Phase III were presented by dimension to a group of 30 sales people, eight district sales managers, and one regional manager. In a four-hour workshop session, these participants independently rated the specific descriptors on a scale with intervals of .25, ranging from .00 (low performance) to 2.00 (high performance).

Means and standard deviations were computed for the ratings for each specific descriptor by the sales personnel and managers. Selection of specific descriptors to be used as the final *BARS* after this phase of the program was made

on the basis of: (1) high agreement between sales people and managers; (2) coverage on the nine-interval scale (.00 to 2.00); and (3) small standard deviations.

Phase V. The eight regional sales managers and fourteen district sales managers who were to participate in Phase VI were given a brief four-hour training session. None of these managers had participated in Phases I through IV. The session was designed to familarize the raters with the form to be used, how it was developed, and various performance appraisal pitfalls (Latham, Wexley, & Pursell, 1975).

Phase VI. The final *BARS* form was used by the regional sales managers ($N = 8$), and the district sales managers ($N = 14$) to evaluate 132 sales personnel. For each dimension, correlations were obtained between two sets of evaluation dimensions—one provided by the regional managers and one provided by the district managers—for those personnel who were common to both samples.

Analysis

A multitrait-multirater analysis (Campbell & Fiske, 1959; Lawler, 1967) was undertaken to examine the validity of the sales personnel *BARS*. A matrix of intercorrelations between performance dimensions was analyzed in terms of convergent and discriminant validity. Convergent validity was reflected in agreement among the regional and district sales managers in assessing the dimensions. This was reflected by correlations between the same performance dimensions, as rated by regional and district managers, being significantly different from zero.

Discriminant validity was indicated by the independence of the performance dimensions. Campbell and Fiske (1959) propose three criteria for demonstrating discriminant validity. First, in the context of this study discriminant validity would exist if the correlation between raters for a dimension was higher than the correlation between that dimension and any other dimension which has neither dimension nor rater in common. Second, a dimension should correlate more highly with an independent effort to measure the same dimension than with measures designed to assess different dimensions which employ the same rater. Third, the same patterns of dimension intercorrelations should exist for all common and different rater combinations.

It is possible for ratings to have convergent and discriminant validity and still not be what would be called *valid* measures of the dimension to be measured. For example, superior and supersuperior ratings of sales personnel may agree perfectly so that convergent and discriminant validity are obtained. However, both superiors and supersuperiors may be making the same incorrect inferences about sales personnel behavior. This would suggest that despite the existence of convergent and discriminant validity, the ratings would be invalid. This kind of situation is probably rare. Still, this possibility clearly illustrates that it is impossible to ever finally validate a set of criteria used for rating individuals.

Results

The Phase I group initially identified 19 sales job performance dimensions, but reduced the number to six by the end of the second day. The job performance dimensions along with their general descriptors were examined by the Phase II participants and only minor alterations in the descriptor terminology were made. The six dimensions and the high, average, and low general descriptors are presented in Table 1. Phase II participants developed 21 to 42 specific descriptors (critical incidents) for each of the six sales job performance dimensions.

The results of the process used in Phase III are summarized in Table 2. If a specific descriptor did not meet the 70 percent retention criterion, it was carefully reassessed to ascertain the reasons for not being retained.

Table 1
Sales Job Performance Dimensions and General
Descriptors (High, Average, Low): Phase I and Phase II

Performance Dimensions	
1. Planning Expertise	
	Does a superior job in planning calls, preparing route schedules, sales presentations, and knowing what competition is doing.[a]
	Is usually well prepared for visiting customers and makes an effort to learn about competition.[b]
	Has difficulty keeping call schedule and often misses scheduled meetings with customers. Is not able to react to customers questions about competition because no effort exerted to learn about competition.[c]
2. Cooperative	
	Follows carefully and promptly requests from superiors for formal information by preparing accurate field reports and by providing valuable unsolicited information to superiors about competitors, customers, and potential customers.
	Makes an effort to be prompt and accurate in submitting important field reports.
	Typically must be prodded to complete necessary paperwork.
3. Independence	
	Is able to make sound sales decisions without always requesting advice from peers or superiors. Understands clearly when advice and counsel should be sought.
	Has the ability to initiate and close most routine sales without requesting help or checking on every detail.
	Is not capable of making a decision on routine or complex tasks without asking for help and advice.
4. Orders/Customers	
	Has an ability to generate orders from different classes of customers.
	Is able to generate orders from a limited class of customers.
	Is not able to consistently generate orders from any particular class of customers.

Table 1 (Continued)

5. Market Coverage

Works the assigned territory thoroughly and is able to establish rapport with all classes of customers in the territory.
Is able to cover most of the territory and has gained the respect of most customers.
Fails to cover the assigned market and is not knowledgeable about some of the customers in the territory.

6. Direct Cost

Is able to generate excellent sales without exceeding reasonable budgeted expense amount.
Can generate acceptable levels of sales within the budgeted expense amount.
Is often faced with the dilemma of exceeding expense budgets even when only generating acceptable levels of sales.

---

[a] The first statement for each dimension is the high performance general descriptor.
[b] The second statement for each dimension is the average performance general descriptor.
[c] The third statement for each dimension is the low performance general descriptor.

Table 2
Agreement on Assigning Specific Descriptors to Six Sales Job Performance Dimensions

Performance Indicators	Original Number of Items	Number of Items Meeting* 70% Retention Criterion
1. Planning Expertise	42	33
2. Cooperative	38	32
3. Independence	40	31
4. Orders/Customers	24	21
5. Market Coverage	21	18
6. Direct Cost	23	19

* Items retained if 25 or approximately 70% of the Phase III participants (30 sales people and 6 district managers).

Convergent Validity Results

Test 1. Table 3 presents the multitrait (dimensions)—multimethod (regional and district managers) analysis (Campbell & Fiske, 1959; Lawler, 1967) derived from assessing the sales occupation-specific *BARS*. The convergent validity test (see Significance Test 1 in Table 3) indicates that all six of the convergent validity coefficients were significantly different from zero. The validities were moderate, ranging from .49 to .39. Note also the moderate inter-rater reliabilities shown in the main diagonal in Table 3. They ranged from .63 to .77, which can be interpreted as moderate rater knowledge of the ratees for district managers. The reliabilities ranged from .48 to .61 for the regional managers, which was somewhat lower than the district manager range.

Of the 30 correlations in the monorater (solid line) triangles, only six were not statistically significant. Of the thirty correlations in the heterotrait-heteromethod triangles, a total of ten were not statistically significant. In order to test more thoroughly and stringently for discriminant validity, the tests recommended by Campbell and Fiske (1959) and House and Rizzo (1972) were conducted. The results are indicated in Tests 2 through 4 of Table 3.

Test 2. The heterotrait-heteromethod correlations are in the dotted line matrix. This test requires that any diagonal entry (convergent validity) should be greater than coefficients in its own row or column. All six of the performance dimensions failed to meet this criterion at the .01 level.

Test 3. This test requires that convergent validities should be greater than coefficients in the equivalent column and row in the upper solid line triangular matrix. Once again, as in Test 2, all six of the performance dimensions failed to meet the criterion at the .01 level.

Test 4. If one group of raters exhibits a halo pattern, then a similar halo pattern should be present in the ratings of the other rater groups, and also between the two groups. A test of this criterion of discriminant validity was accomplished by ranking the coefficients within the

Table 3
Validity Correlation Matrix: District and
Regional Managers As Raters Using the Bars for 102 Sales People

Performance Dimensions	District Sales Managers						Regional Sales Managers					
	PL	CO	IN	O/C	MC	DC	PL	CO	IN	O/C	MC	DC
PL	(71)											
CO	51	(73)										
IN	42	-38	(65)									
O/C	64	43	09*	(69)								
MC	53	-31	-29*	-06*	(63)							
DC	55	64	56	10	58	(77)						
PL	(49)	53	-06*	-09*	38	37	(49)					
CO	38	(48)	18	-10*	21*	-08*	81	(52)				
IN	53	47	(39)	37	19*	29	63	49	(61)			
O/C	-15*	49	49	(47)	17	41	10*	68	37	(48)		
MC	-07*	53	37	58	(42)	40	71	49	62	19*	(61)	
DC	46	46	31	-09*	-02*	(45)	58	58	12*	48	59	(51)

Significance Tests

	PL	CO	IN	O/C	MC	DC
1. Significance of validity coefficients (r_{ii}'s in parentheses)	<.01 (.01)	<.05 (.05)	<.01 (.01)	<.01 (.01)	<.01 (.01)	<.01 (.01)
No. of heterotrait hetermethod coefficients > validity coefficients	2	3	3	3	2	2
No. of ties	0	0	0	0	0	0
2. p (sign test); n = 10-ties	.055	.172	.172	.254	.055	.055
No. of monomethod-heterotrait coefficients > validity coefficients	4	2	2	2	2	4
No. of ties	0	0	0	0	0	0
3. p(sign test); n = 5-ties	.969	.500	.500	.500	.500	.969

4. Coefficient of Concordance W = .15; p < .20

* All coefficients with asterisks were not significant at the p < .05 level. All other coefficients were at least significant at the p < .05 level.

Abbreviations: PL = Planning Expertise; CO = Cooperative; IN = Independence; O/C = Orders, Customers; MC = Market Coverage; DC = Direct Cost.

four Table 3 triangles and computing a coefficient of concordance, W, to ascertain whether there was a significant similarity among the patterns of correlation coefficients. The coefficient of concordance was not statistically significant at the .01 level.

Discussion

An issue of interest in the present study is whether previous *BARS* research could be generalized to an occupational group such as sales personnel. The application of the multitrait-multirater analysis revealed moderate evidence of convergent validity of the performance dimensions. However, little evidence of discriminant validity was found. These findings are similar to previous multitrait-multimethod results involving nurses and firefighters.

Convergent Validity

The results shown in Table 3 reveal moderate correlations in the marked-off diagonal section of the matrix. This suggests that there is some agreement between the two groups of raters about the *BARS* traits. Overall, however, the convergent validity would be considered moderate instead of strong.

There appear to be a number of plausible explanations for the moderate degree of convergent validity obtained in the present study. First, the district sales managers and regional sales managers are two supervisory levels, each with different functions. Second, the district sales managers and regional sales managers do not have equivalent opportunities or time to observe and evaluate the sales people. For example, as mentioned, the district managers occasionally observe a sales person selling the company product line when a difficult sale is being attempted. The regional managers interact less frequently with sales personnel. Third, it is not known whether the district sales managers and regional managers have the same type of expectations or values concerning performance. If different expectations and/or values exist, moderate-to-weak convergent validity could be the result.

Discriminant Validity

In the present study, discriminant validity was not found for any of the significance tests conducted. It was assumed that the development of a *BARS* through superior-subordinate participation would yield independent sales job performance dimensions. However, as Table 3 illustrates, there are high interdimensional correlations. The failure of the *BARS* to meet the stringent discriminant validity tests could be the result of halo and method bias.

A second possibility for the lack of discriminant validity is that raters (managers) had inadequate job-relevant contacts with ratees (sales personnel). The district and regional managers would have a difficult time using the *BARS* as a basis for discriminating among ratees without increasing their observations of the job performance of sales personnel.

The discriminant validity criteria proposed by Campbell and Fiske (1959) are very rigorous. Each test is quite stringent and would be difficult to meet. In fact, there are few studies reported in the literature involving performance appraisal ratings that meet each of the discriminant validity criteria.

The results of this study provide management with objective data concerning the use of raters who are interpersonally and geographically removed from ratees. The regional managers clearly indicate a halo influence in their *BARS* ratings. Perhaps a more valid final *BARS* would have evolved if more regional managers had participated in each of the phases in the study. The limited number of regional managers, however, did not permit this in the present study.

The need for more research comparing *BARS* and other types of performance appraisal systems is obvious (Bernardin, Alvares, & Cranny, 1976). Theoretical advantages, moderate convergent validity, little discriminant validity, and moderate reliability in themselves do not justify the implementation of a *BARS* system for performance evaluation, the identification of training needs, sales personnel selection decisions, or other similar sales-oriented projects. Even if strong convergent and discriminant validity is

found in a study of a performance appraisal system such as *BARS*, there is the possibility that the raters' ratings are invalid.

Despite its low validity *BARS* might be useful for the sales personnel employed in the company studied, although rigorous testing of validity is needed before applying a performance appraisal program. The *BARS* in this study provided raters with counseling and feedback information. However, the discriminant validity was not sufficient to state that the sales-personnel *BARS* was psychometrically sound. Perhaps more training in the development of the *BARS* itself coupled with its actual use could improve both the convergent and discriminant validities. Certainly, more research on the *BARS* versus other traditional scales is warranted and could improve our knowledge of this method of appraising performance.

References

Arvey, R. D., & Neel, C. W. Testing expectancy theory predictions using behaviorally-anchored based measures of motivational effort for engineers. *Journal of Vocational Behavior,* 1974, *4,* 299–310.

Bernardin, H. J., & Walter, C. S. Effects of rater training and diary keeping on psychometric error in ratings. *Journal of Applied Psychology,* 1977, *62,* 64–69.

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

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81–105.

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

Campbell, J. P., Dunnette, M. C., Lawler, E. E., & Weick, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Dickinson, T. L., & Tice, T. E. A multitrait-multimethod analysis of scales developed by retranslation. *Organizational Behavior and Human Performance,* 1973, *9,* 421–438.

Fogli, L., Hulin, L. C., & Blood, M. R. Development of first-level behavioral job criteria. *Journal of Applied Psychology,* 1971, *55,* 3–8.

Goodale, J. G., & Burke, R. J. Behaviorally based rating scales need not be job specific. *Journal of Applied Psychology,* 1975, *60,* 389–391.

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

House, R. J., & Rizzo, J. R. Toward the measurement of organizational practices. *Journal of Applied Psychology,* 1972, *56,* 388–396.

Landy, F. J., & Guion, R. M. Development of scales of the measurement of work motivation. *Organizational Behavior and Human Performance,* 1970, *5,* 193–203.

Latham, G. P., Wexley, K. N., & Pursell, E. D. Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology,* 1975, *60,* 1550–1555.

Lawler, E. E., III. The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology,* 1967, *51,* 369–381.

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

Williams, W. E., & Seiler, D. A. Supervisor and subordinate participation in the development of behaviorally-anchored rating scales. *Journal of Industrial and Organizational Psychology,* 1973, *4,* 1–12.

Zedeck, S., & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance,* 1972, *7,* 457–466.

Author's Address

Joseph M. Ivancevich, Professor of Organizational Behavior and Management, University of Houston, TX 77004