Journal of Applied Psychology
1981, Vol. 66, No. 4, 458-463

0021-9010/81/6604-0458$00.75

A Clarification of Some Issues Regarding the Development and Use of Behaviorally Anchored Rating Scales (BARS)

H. John Bernardin
Virginia Polytechnic Institute and State University

Patricia Cain Smith
Bowling Green State University

This article attempts to clarify some issues regarding the development and use of behaviorally anchored rating scales (BARS). The Smith and Kendall (SK) format is distinguished from other approaches to BARS, and research comparing their relative effectiveness is reviewed. The SK format is discussed as a method that is designed to enhance future observations and to foster a common frame of reference in observer raters. Emphasis is also placed on BARS as an observation-rating system that provides data for the assessment of estimates of accuracy for individual raters. Responses are made to criticisms dealing with the rating process of BARS, the relative effectiveness of BARS versus summated scales, and the role and characteristics of the behavioral anchors. Further research is recommended on the development and use of BARS.

Some time ago, the first author of this article submitted a manuscript to one of the applied journals concerning rating approaches for behaviorally anchored rating scales (BARS). One of the reviewers kindly asked, "How long are we going to dance on the head of the BARS pin?" and concluded that we already know that BARS are certainly no better and probably worse than other rating formats. Some recently published articles have generally supported the latter position (e.g., Atkin & Conlon, 1978; Latham, Fay, & Saari, 1979). These articles have attacked the theoretical, conceptual, practical, and empirical justification for the use of BARS. The purpose of this article is to clarify and reemphasize several elements of the BARS approach that distinguish it from other methods and that have seemingly been ignored or misinterpreted. Responses will be made to the major criticisms of the approach, and recommendations will be made with respect to future usage and research.

A version of this article was presented at the annual meeting of the Academy of Management, Detroit, Michigan, August 1980. Support for this research was partially provided by National Institute of Education (NIE), Department of Education, Grant 110-053-363933-11, H. John Bernardin, principal investigator. Views in this article do not necessarily reflect those of NIE or of any other agency of the U.S. government.

The authors wish to thank Liesa Bernardin for assistance in the preparation of this manuscript.

Requests for reprints should be sent to H. John Bernardin, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

The Smith and Kendall Approach to BARS

Smith and Kendall introduced BARS in 1963 in a National League for Nursing study (Smith & Kendall, 1963). The procedure was designed to encourage raters to observe behavior more carefully, to infer the meaning of that behavior, and to record observed incidents on a continuum of effectiveness for specific dimensions. It was hoped that the method would allow for a record of observation that was sufficiently clear so that it could be discussed with the person who was rated and so that a summary rating could more easily be made at a later date, if necessary. The emphasis by subsequent researchers and writers on the BARS approach has apparently disregarded the important issues of observation and interpretation. Rather, the focus has been on the summary rating aspect of the process. In addition, there has been little consideration of BARS as a method for enhancing the validity of future observations. Put another

way, BARS was designed to standardize not only the rating process but also the observation process (Jacobs, Kafry & Zedeck, 1980). There has been very little comment on this equally important aspect of the appraisal process.

When people make summary ratings based on retrospection after periods as long as a full year, their judgments are often biased. The "gating" process screens out observations that are not in agreement with prior stereotypes (Cooper, in press). Moreover, broad generalizations are made from a few, usually recent, observations (if, indeed, any observations are made at all). We believe the evidence is strong that the typical judge is not cognitively prepared to summarize and abstract adequately from a great many observations if no record of observations has been maintained throughout the appraisal period. The essence of the BARS approach as designed by Smith and Kendall (1963) was to enhance and standardize observation and thus prepare raters for summary ratings that may be necessary in the future. Borman (1979) discussed the need to develop a common frame of reference for the observation of behavior. The use of BARS throughout an appraisal period should reduce the idiosyncrasies in raters' perception as they observe the same or similar ratee behavior. Thus the idea was to foster the development of valid stereotypes of effective and ineffective performance prior to observation.

The Smith and Kendall (SK) format must be distinguished from the wide variety of forms and methods grouped under the rubric of BARS (e.g., Atkin & Conlon, 1978; Jacobs et al., 1980). The SK format was derived from a mixture of the Fels Parent Behavior Rating Scales (Guilford, 1954, pp. 266–267) and Thurstone's attitude scales (Guilford, 1954, pp. 456–459). With the SK format, raters were to be given a set of vertical graphic scales and instructed to record the behavior observed on each applicable scale throughout the appraisal period. The instructions were to observe the behavior, decide to which dimension it belonged, and then indicate on the scale the date and details of the incident. The notation of the incident was to be made at the effectiveness level on the scale that was considered to be the most appropriate for that incident on that behavioral dimension. The scaling of the effectiveness level of the observation, that is, the place on the page at which the observer recorded the incident, was to be aided by a comparison with a series of illustrative behavioral "anchors" and generic descriptors. The illustrative behavior had been previously scaled as belonging to a particular dimension and as representing different effectiveness levels for that dimension. Three more dimension clarification statements defined the highest, lowest, and mid-point of each scale. It was not necessary that the notation of observed behavior be made at the exact point on the graphic scale at which some illustrative behavior had been previously scaled. Rather, the observer was supposed to infer the behavioral dimension involved and then scale what had been observed in relation to the specific and more generic examples. The rater would thus interpolate between the illustrative examples when recording a brief notation of the behavior that had been observed.

The observer would be scaling by direct estimation of magnitude, presumably a prothetic scale (Stevens & Galanter, 1957). The anchoring illustrations were to be concrete and specific and located at irregular intervals along the relevant scale based on their scaled effectiveness levels. After a period of observation and incident recording, the rater could, if necessary, make a summary rating. This summary, in addition to the notes, could serve as a basis for discussion with the person who was observed and/or as a criterion measure.

Thus, to summarize the SK format, the sequence was to be as follows: observation—inference—scaling—recording—summary rating. The process sought to define, clarify, and operationalize the implicit evaluative theory of the rater. The desired psychological process of the rater dictated the choice of graphic scales. The purpose was to encourage observation and explicit formulation of the implications and interpretations of behavior. It was really the emphasis on the approach as a method for enhancing future observations that distinguished it from other approaches such as forced choice, summated, and simple graphic scales. The ap-

proach was designed to facilitate a common frame of reference in observers so that they would look for the same kind of behavior and interpret them in essentially the same way. In addition, the sequence provided an opportunity to verify and validate summary ratings by individual raters that were made after lengthy performance periods (see Bernardin, 1979).

Major Criticisms of BARS

There have been a number of issues raised recently regarding the development and use of BARS. The most important criticisms have dealt with the rating process used with BARS, the relative effectiveness of BARS versus summated scales, and the role and characteristics of the behavioral anchors. Although some of these criticisms are justified and underscore the need to improve instrumentation, others require a response.

The BARS Rating Process

There have been several discussions of BARS as strictly a rating format in which the rater (a) reads the definition of the performance dimension (b) reads each critical incident in some prescribed order; and (c) marks the incident that represents the "most typically expected" behavior. The scale value of this incident is then to be taken as the rating on that dimension (e.g., Atkin & Conlon, 1978; Latham et al., 1979). This discription is at variance with the approach discussed by Smith and Kendall (1963) and delineated above. The SK approach recommended that the incidents be used to define the continuum of performance on a given dimension along with the dimension clarification statements. If the incidents were to be used as suggested above, where only one is selected per dimension as "most typical," there would be no reason to retain them within a graphic format. Rather, as Smith and Kendall stated, the incidents were included to give the rater an idea of where different types of behavior are rated in the context of the dimension definition and the elaboration statements at the top, mid-point, and bottom of each scale. All of this information was to be used to operationalize the continuum of performance in order to standardize (as much as possible) both the observation and the rating processes. Merely reading the examples in ascending and descending order as others have suggested, as if determining a sensory threshold, is probably justified only for observations made over a very short time span, such as an interview (Maas, 1965).

As stated above, Smith and Kendall (1963) also called for the observer/rater to write and scale observed behavior on each dimension for each ratee. This approach has been used in several studies (Bernardin, 1977; Bernardin & Walter, 1977; Tate, 1964; Zedeck & Baker, 1972; Zedeck, Kafry, and Jacobs, 1976; Hom, DeNisi, Kinicki, & Bannister, Note 1). Bernardin, LaShells, Smith, and Alvares (1976) compared this approach with one that is very similar to that described by Atkin and Conlon (1978). Using the mean of the newly scaled items written for each dimension as the rating that was compared with the "checked" rating, Bernardin, LaShells et al found lower leniency effect, increased rating variability across dimensions, and greater discriminability in ratings when students wrote and scaled incidents than when they simply checked a point on the scale. Bernardin (1977) also found substantial improvement in the psychometric quality of ratings from BARS when the scaled incident procedure was used instead of the checking procedure (Bernardin, Alvares, & Cranny, 1976).

One obvious criticism of this approach is the time involved in writing and scaling incidents for each dimension. Although this is a legitimate criticism, we would argue first that the amount of time required for writing and scaling incidents on BARS is well worth it, given the need to document and justify numerical ratings. Second, the time required for recording observations on several ratees may not be much greater than the time required to rate the same number of ratees on a summated scale composed of 40–100 items. Given a normal span of control where a supervisor must evaluate six people, with a 50-item summated scale, the rater must make 300 ratings per appraisal period.

The Role and Characteristics of the Behavioral Anchors

Atkin and Conlon (1978) made several references to a neutral point on the scales, above which a good employee would always be rated and below which a bad employee would always be rated. With this exception, at no place in the BARS literature can reference be found to this neutral point. Although there will always be a mid-point for each scale and each person with a scaled incident in close proximity to it, the incident could conceivably represent good behavior, bad behavior, or probably average behavior (as defined in the clarification statement). Item traces along the behavioral continuum will rise to a peak and then fall in approximately a normal distribution.

Contrary to statements by several writers, the SK format was not intended to be cumulative in the Guttman sense. Observed behavior of a single ratee was expected to cluster reasonably close to a point on the scale, but the format was designed to allow for intraindividual variability across an appraisal period within a given behavioral domain. The simple checking procedure described above precludes a consideration of this potentially valuable source of variance. In their discussion of criteria for appraisal effectiveness, Kane and Lawler (1979) stated that an appraisal system preoccupied with the measurement of ratees' "typical" performance (such as a BARS checking procedure) ignores "the massive accumulation of evidence that performance . . . is determined at least as much by variable intra- and extra- individual factors as by traits" (p. 433). The SK format calls for the generation of a performance distribution for each ratee throughout the appraisal period. Thus differences in what Thorndike (1949) called intrinsic unreliability can be measured for individuals.

Another problem with the checking procedure in which the rater selects the "most typical incident" has to do with the distribution of perceived effectiveness for each anchor. Regardless of the minimum requirements set for the standard deviations of item effectiveness ratings, distributions of incidents that are adjacent on a scale will overlap in their perceived effectiveness. In fact, in a format such as that described by Atkin and Conlon (1978) with nine incidents corresponding to nine scale points, the probability would be great that at least one pair of incidents will "flip-flop." Such a situation creates havoc for the rater who receives directions for checking the "most typically expected" incident. A justifiable criticism of BARS is that even given low standard deviations for the incidents, it is conceivable that a particular ratee could have behaved precisely as described in an item that was scaled higher in effectiveness and also precisely as described in an item that was scaled below it. The use of the index of reproducibility or scalogram analysis to screen items increases the chances for the "monotonicity" of the items across all raters. Although reproducible items in a Guttman sense are certainly desirable, the items were included only to serve as conceptual anchors for future observation and subsequent rating. As stated earlier, the SK format called for a summary rating in the future based on a consideration of the numerous newly scaled observed behaviors recorded throughout the appraisal period and scaled with the established anchors as a context.

Borman (1979) discussed another problem regarding the behavioral anchors on BARS. Raters, he stated, "often have difficulty discerning any behavioral similarity between a ratee's performance and the highly specific behavioral examples used to anchor the scale" (p. 412). This is a legitimate criticism and particularly troublesome when raters are instructed simply to select the "most typical incident" representative of the ratee. Smith and Kendall (1963) discussed this problem, stating that "a specific critical behavior could not occur and hence could not serve as a basis for rating" (p. 150). This, however, is the reason why the SK format called for the use of more generic statements to anchor the scales in addition to the specific behavior. Bernardin, LaShells et al. (1976) found that the inclusion of these generic statements significantly improved the quality of the ratings. Also, as stated above, the rating process would be augmented by

the observations made by the raters throughout the appraisal process.

BARS Versus Summated Scales

Atkin and Conlon (1978) recommended that a new format be adopted in which the rater would judge "whether each incident on every dimension would or would not be expected" of the employee being evaluated. This approach appears to be essentially the same as summated scales, introduced as behavioral observation scales by Latham et al. (1979). Although comparisons between BARS and summated scales with regard to rating errors and reliability have not strongly supported either method (e.g., Bernardin, 1977; Hom et al., Note 1), one problem that is characteristic of summated scales may have relevance to their potential validity. Kane and Bernardin (Note 2) state that summated scales, unlike BARS, fail to account for the fact that an occurrence rate may represent a different degree of effectiveness depending on the particular form of behavior to which it applies. For example, in a police detective's job, an 85%–94% occurrence rate may constitute superior performance in obtaining arrest warrants within 3 months in homicide cases but horrendous performance in being vindicated by the Internal Review Board in instances of having used lethal force. This use of fixed standard scaling that is characteristic of summated scales can result in serious errors in the levels of "satisfactoriness" to which ratees are assigned on each dimension. Parenthetically, the problem may account for the fact that in the only published review of Thurstone versus summated scales, Seiler and Hough (1970) found Likert scales to be about 10 points higher than Thurstone scales in reliability, but the validities were about equal. Thus, with corrections for attenuation, validities could be higher for Thurstone scales.

Conclusion

Atkin and Conlon (1978) concluded that we need a break from our "myopic concern with instrumentation" (p. 128) and a reconcentration of effort on the rater and the rating process. It is evident to us, however, that the problem may be more hypermetropic in nature. It may be that the variety of rating formats and developmental procedures used under the guise of BARS at least partially accounts for the disappointing results (Jacobs et al., 1980).

Although it is true that there has been an abundance of empirical comparisons between BARS and other formats, BARS format and rating procedures have not been subjected to the same methodological scrutiny that followed, for example, the introduction of summated or Guttman scales. Furthermore, as delineated above, even the few empirically supported optimal rating and developmental procedures have often been ignored. The discussion of BARS in numerous articles and the development and use of BARS in many studies are analogous to the use of summated rating scales that have not been item analyzed and that have a 2- or 15-point response format. In short, there are data that show that instrumentation does make a difference.

It is not our intention to imply that the SK approach to BARS is the "truth" in appraisal and cannot be improved. We believe, however, that changes should be made thoughtfully. Since the introduction of the procedure in 1963, there have been few attempts to scrutinize BARS methodology (Jacobs et al., 1980). More studies of this nature are needed to improve instrumentation.

It is clear from the literature that one cannot develop BARS in a haphazard fashion or use any format she or he feels might work and expect positive results every time. As for virtually every other rating system before it, there are proposed and tested methods for ensuring maximum efficiency from BARS. Researchers would do well to adhere to these methods, improve them, and test others before concluding either that instrumentation makes little difference or that BARS are not worth the time or effort. Several writers have discussed the qualitative advantages of BARS (e.g., Jacobs et al., 1980). We believe that quantitative advantages may also be manifest, given a stricter adherence to empirically supported methodology.

It is also clear from the literature that asking a person to do a summary rating after

6 to 12 months of performance without any record of observation invites virtually every type of rating error possible. Smith and Kendall (1963) stressed this point in their call for the maintenance of a record of observed behavior and scaling of that behavior in the context of the conceptual anchors, the dimension clarification statements, and the dimension definitions. This procedure is important not only as a basis for verifying any future summary ratings but also for enhancing future observations and standardizing the observational process. We believe this to be a major factor that distinguished BARS from other methods as an observation-rating system. The purpose of this article was to clarify some of the methodology proposed by Smith and Kendall and to comment on some of the criticisms that have been leveled against BARS. We have attempted to show that most of these criticisms are actually directed at BARS methods that are quite different from the Smith and Kendall approach.

Although, as numerous writers have suggested, greater sources of variance can probably be found in rater characteristics and organizational context, methodology is a source of variance that psychometricians can control. We, therefore, would do well to pay even greater attention to the essentials of instrumentation.

Reference Notes

1. Hom, P. W., DeNisi, A. S., Kinicki, A. J., & Bannister, B. D. *Behaviorally anchored rating scales versus summated scales: Psychometric properties, resistance to bias, and feedback effectiveness.* Unpublished manuscript, 1981. (Available from P. W. Hom, School of Business, Kent State University, Kent, Ohio 44240.)
2. Kane, J. S., & Bernardin, H. J. *Further reflections on the development and justification of the BOS methodology.* Unpublished manuscript, 1980. (Available from J. S. Kane, Office of Personnel Management, Washington, D.C. 20415.)

References

Atkin, R. S., & Conlon, D. J. Behaviorally anchored rating scales: Some theoretical issues. *Academy of Management Review*, 1978, *3*, 119–128.

Bernardin, H. J. Behavioral expectation scales versus summated scales: A fairer comparison. *Journal of Applied Psychology*, 1977, *62*, 422–427.

Bernardin, H. J. Implications of the Uniform Guidelines on Employee Selection Procedures for the performance appraisal of police officers. In D. C. Spielberger (Ed.), *Proceedings of the National Workshop on the Selection of Law Enforcement Officers*. Tampa, Fla.: Human Resources Institute, 1979.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. A re-comparison of behavioral expectation scales to summated scales. *Journal of Applied Psychology*, 1976, *61*, 284–291.

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. Behavioral expectation scales: Effects of developmental procedures and formats. *Journal of Applied Psychology*, 1976, *61*, 75–79.

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

Borman, W. C. Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, 1979, *64*, 410–421.

Cooper, W. H. Ubiquitous halo. *Psychological Bulletin*, in press.

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

Jacobs, R., Kafry, D., & Zedeck, S. Expectations of behaviorally anchored rating scales. *Personnel Psychology*, 1980, *33*, 595–640.

Kane, J. S., & Lawler, E. E., III. Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), *Research in organizational behavior* (Vol. 1). Greenwich, Conn.: JAI Press, 1979.

Latham, G. P., Fay, C., & Saari, L. M. The development of behavioral observation scales for appraising the performance of foremen. *Personnel Psychology*, 1979, *32*, 299–311.

Maas, J. B. The patterned scaled expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology*, 1965, *49*, 431–433.

Seiler, L. H., & Hough, R. L. Empirical comparisons of the Thurstone and Likert techniques. In G. F. Summers (Ed.), *Attitude measurement*. Chicago: Rand-McNally, 1970.

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.

Stevens, S. S., & Galanter, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, *54*, 337–411.

Tate, B. *Test of a nursing performance evaluation instrument*. New York: National League for Nursing, 1964.

Thorndike, R. L. *Personnel selection: Test and measurement techniques*. New York: Wiley, 1949.

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.

Zedeck, S., Kafry, D., & Jacobs, R. Format and scoring variations in behavioral expectation evaluations. *Organizational Behavior and Human Performance*. 1976, *17*, 171–184.

Received August 28, 1980 ■