

CONSTRUCTION OF BEHAVIOURALLY ANCHORED RATING SCALES (BARS) FOR THE MEASUREMENT OF MANAGERIAL PERFORMANCE

H.H. SPANGENBERG

J.J. ESTERHUYSE

J.H. VISSER

J.E. BRIEDENHANN

*Stellenbosch Farmers' Winery
Stellenbosch*

C.J. CALITZ

*Department of Industrial Psychology
University of Stellenbosch*

ABSTRACT

BARS were initially developed as indices of behavioural change and to ensure greater comparability of ratings from different raters. In this study, BARS were developed for a major producer-wholesaler company in the liquor industry to serve as an independent criterion in the validation of the company's assessment centre, to assess the impact of development activities on the skill levels of assessment centre participants and as a diagnostic tool in identifying performance deficiencies. A step-by-step account of the four stages in the development of BARS is presented, together with examples of actual scales for the final steps.

OPSOMMING

Gedragseankerde skale (BARS) is oorspronklik ontwikkel as indekse van verandering, en om die vergelykbaarheid tussen beoordeelings van verskillende beoordeelaars te verhoog. In hierdie studie is BARS vir 'n groothandelaar in die drankbedryf ontwikkel ten einde te dien as 'n onafhanklike kriterium in die validering van hulle takseersentrum; om die invloed van ontwikkelingsaktiwiteite op die vaardigheidsvlakke van deelnemers aan die takseersentrum te meet; en as 'n diagnostiese hulpmiddel in die identifisering van ontoereikende prestasie. 'n Stap-vir-stap beskrywing van die vier stadia in die ontwikkeling van BARS word gegee, met voorbeelde van werklike skale vir die finale stappe.

Behaviourally Anchored Rating Scales (BARS) were first developed by Smith and Kendall in 1963 and initially called "Unambiguous Anchors for Rating Scales". Their rationale for the development of these scales was their concern about the comparability of ratings that were normally used for validation of tests and as indices of effectiveness of educational, motivational and situational changes. They argued that ratings from different raters in different situations should in fact be equal since they are almost always treated as if they were. Furthermore, this demand for comparability meant that interpretation of the rating must not deviate too widely from rater to rater or from occasion to occasion, either in level (evaluation) or in dimension (trait, situational characteristics, job demands, etc.) (Smith & Kendall, 1963, p. 1).

The format they proposed for rating scales and which is still used today, with some variations, is a graphic rating scale, arranged vertically. For each dimension, behaviour examples typifying various points on the

scale are printed (see Figure 1 for a typical example of a scale).

Langford (1980), in a comprehensive article on managerial effectiveness criteria, describes the basic method for the development of BARS as follows:

"The method they (Smith & Kendall) used was based on the retranslation technique used in languages; for example, a piece of English is translated into another language by one person, retranslated back by another into original English and the two compared for accuracy. The method involves five steps, which are iterative:

1. Generation of qualities/dimensions, i.e. which aspects of job behaviour should be evaluated, by a group of judges with experience of the job in question.
2. Generation of behavioural statements representing effective, ineffective and average performance for the job in question by the same group of judges. Editing of these statements into 'expectations' of a specific behaviour by adding the prefix 'could be expected to . . .'
3. Allocation of statements to dimensions, usually by the researchers.

4. Reallocation (or retranslation) of the statements to dimensions, by a separate but comparable group of judges. Statements and dimensions are retained or rejected according to a previously established criterion percentage.
5. Scaling of statements for each dimension on a rating scale of between 5 and 9 points ranging from very ineffective to very effective. Retention or rejection of statements according to level of dispersion and standard deviation" (pp. 100 – 101).

According to Langford, the antecedents of BARS can

be found in Flanagan's Critical Incident Technique and the work of Bendig, who found that the reliability of rating scales could be increased by using verbal anchors. Although this method has since been applied to diverse areas of measurement (eg. training, selection, and motivation), very few studies have been undertaken in which BARS were used to measure managerial effectiveness. The only research that could be found in this regard was the unpublished work done by Slivinski and his co-workers at the Canadian Public Service, which is highly significant for the innovations it generated in the field of assessment centres (Slivinski, Grant, Bourgeois & Peterson, 1977).

SELF-CONFIDENCE

He is realistic and has a positive self image; acts confidently in a variety of situations.

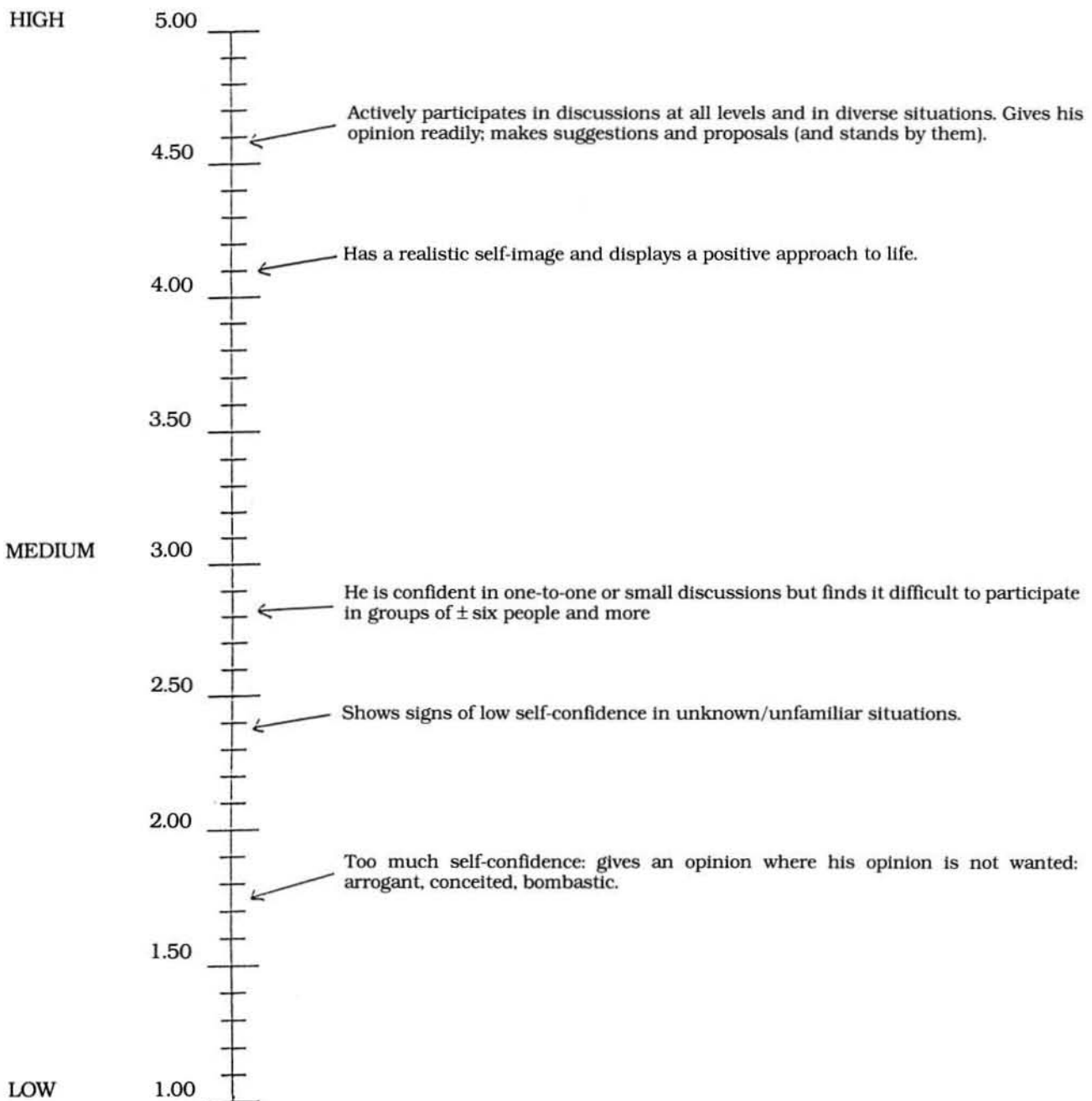


Figure 1. Example of a Behaviourally Anchored Rating Scale

In their Customs and Excise Assessment Centre for *first line supervisors*, they used the process of development of BARS for the development of rating scales to measure performance. Considering that the identification of dimensions is an integral aspect of the development of BARS, their approach is very meaningful.

Slivinski et. al. (1977, p. 30) describes the advantages of the BARS system as follows:

"First, the scales are rooted in and refer to actual observed behaviours. Secondly, both the dimensions and the behavioural anchors are based on the judgement of experienced managers who understand the nature, functions and responsibilities of the job, and who are reasonably comparable to those who will eventually use the dimensions. In addition, the qualities or characteristics listed are operationally defined in the raters' terminology and are distinguishable from one another by the raters. Finally, the same dimensions and behaviours can be used as criterion measures as well as predictors."

Also significant is their work on the use of BARS in evaluating managerial job performance of participants in an *executive* assessment centre. Because of the complexity and heterogeneity of the executive's job, and based on comprehensive job analyses they modified the basic approach substantially. According to McCloskey (1979), their approach differed from the normal procedure in three important ways:

- The degree of concreteness used in the behavioural anchors (i.e. descriptive anchors representing a *summary* of the many behaviours associated with a particular level of performance as opposed to a *single* specific behavioural example).
- The presentation format (i.e. performance summaries presented randomly, as opposed to behavioural anchors displayed on a graphic scale in a type of sequential order extending from the lowest to the highest levels of performance).
- The procedures used for generating the behavioural summaries (i.e. the summaries were determined by studying assessors' observations of participants' performance in the assessment centre simulation exercises, as opposed to a job analysis approach involving supervisors and/or job incumbents).

These are certainly not only innovative but also drastic deviations from the standard procedure for the development of BARS.

In view of the heterogeneity of the executive position (Mintzberg 1973; Kotter 1982; Luthans & Lockwood, 1984), it seems advisable to cluster behaviours associated with a particular level of performance. The more comprehensive the behaviour clusters, the better the superior should be able to assign ratings accurately.

Presentation of performance summaries in random

rather than sequential order also seems advisable, specifically when use is made of summaries rather than single and specific behaviour examples. Random sequencing does however deny the rater the opportunity to direct his judgement toward a specific area on the scale. Furthermore, evaluators in fact make use of the behavioural anchors and find them very useful. This area definitely requires more research.

Using behaviour examples emanating from assessment centre participation for the construction of BARS is a very exciting innovation. There is, however, a very important prerequisite: that the construction of the entire Centre be based on a job and contextual analysis. If not, the behavioural examples might be of good quality, but irrelevant.

A further contribution of the work by Slivinski and his colleagues was that they used different formats of BARS to suit different levels of management, thereby accommodating the wide differences in complexity between the first level manager and the executive level manager.

METHOD

From the outset there was a short term and a longer term objective for the development of BARS. The immediate one was to use BARS for the validation of a Middle Management Assessment Centre. (Spangenberg, Esterhuysen, Visser, Briedenhann & Calitz, 1988.) The longer term objective was to provide a *measuring scale* directed towards *work behaviour*: a scale that would be based on behavioural examples or incidents that would be relevant to middle and senior managers.

In the long run BARS would be of assistance in the following areas:

- Assessment centre follow-up. One of the problems personnel managers experience in the follow-up process is to determine the influence of development activities on the skills levels of assessment centre participants. Behaviourally based rating scale would assist in supplying that information.
- Performance management. One of the important contributions of performance management to individual performance is that it measures the quality of performance fairly objectively, whether an individual performs adequately or not. It is not so easy, however, to determine the *reasons* for inadequate performance, especially if it is behaviourally related. It is towards the identification of personal or behavioural problems that BARS can make a useful contribution. Where performance management deals to a large extent with the *results* area of performance, BARS can complement it with its emphasis on the *behavioural* side of performance.

During the planning phase for the construction of BARS, careful attention was given to the format of the BARS to be developed. The authors were keen to use the format developed by McCloskey for the reasons described.

Another very important consideration was that, in view of the objective of using BARS for various purposes, i.e. the measurement of on the job performance as well as the effect of assessment centre follow-up, it was important that it should include, if necessary, dimensions which could not be effectively measured in the assessment centre.

Furthermore and equally important, separating the development of BARS completely from that of the assessment centre ensured its independence as a criterion measure.

Step I

The first step in the construction of BARS comprised two phases, i.e., firstly the identification of characteristics required for managerial effectiveness at the target level, (which was the top level of middle management) and secondly, the eliciting of behaviour examples for the dimensions identified. The following procedure was adopted.

In order to procure the necessary information a number of brain storming sessions were held. These sessions were attended by managers from senior levels who, by nature of their positions, had a wide perspective on the content and context of the middle manager's job.

During the sessions participants were asked to identify characteristics or dimensions of managerial effectiveness. Participants studied these dimensions and reduced the number of dimensions by eliminating overlapping concepts.

Another brainstorming session followed during which participants were asked to give critical incidents of behaviour for each of these dimensions. They were asked to give examples of excellent, average and poor behaviour.

Step II

The next step was to process the above information in order to arrive at dimensions with behavioural examples. This was done by the administrators and involved the following activities:

- The information suggested more than thirty dimensions with some degree of overlap between them. It was decided, therefore, to reduce the number of dimensions and to create meaningful constructs. At the conclusion of this exercise 19 dimensions were retained.
- The next step was to “retranslate” behaviour examples elicited during brainstorming sessions to the newly defined dimensions. Behaviour examples which were very similar in content were combined and reformulated. Examples were categorized as high, average or low, as indicated during the brainstorming sessions (Step I).
- Hereafter the coverage of dimensions was studied as well as the extent to which examples of excellent, average or poor behaviour have been elicited for *each individual dimension*. This step is important, since inadequate coverage in whatever

form calls for elimination of a dimension or dimensions. Of the 450 behaviour examples elicited during the brainstorming sessions, 256 reformulated behaviour examples were retained, providing sufficient coverage for each of the 19 dimensions. It was found, however, that relatively more examples were identified at the extremes of the scales than in the middle.

- Finally, for each of the 19 dimensions, a separate page was prepared containing a definition of the dimension as well as the behaviour examples in random order.

Step III

The next step was to assign scale values to the behaviour examples. For this exercise two documents were required:

- A document containing dimensions and behaviour examples (see Table 1).
- A 5-point rating scale for each dimension containing the definition of that dimension.

A group of 25 senior managers was asked to rate each of the behaviour examples. They were required first to read the definition of a specific dimension and then to read each behaviour example and decide on a scale of value for that example, where 5.00 would represent very good behaviour and 1.00 very bad behaviour on the dimension.

Step IV

The last step was to construct the final Behaviourally Anchored Rating Scales. The procedure was as follows:

- From the ratings made by senior managers, the mean and standard deviation were calculated for each behaviour example. For each rating scale, behaviour examples with high standard deviations were not considered. This was done because a high level of dispersion could imply unreliability in the interpretation of a behavior example.
- Finally behaviour examples with mean scores nearest to the high, average and low positions on the scale (i.e. 5, 3 and 1) were identified. The most important criterion for inclusion of a behaviour example was the standard deviation of the specific example. In cases where a behaviour example was content-wise clearly superior to another example which had a marginally higher standard deviation, a qualitative decision was taken as to which example should be included.

As an example, the means and standard deviations of behaviour examples of the Rating Scale for Judgement are presented in Table 2. The behaviour examples chosen for inclusion in the final Rating Scale are indicated as follows: H = high; MH = medium high; M = medium; ML = medium low; and L = low.

Of the 256 behaviour examples available for final construction of the Scales (Step IV), 95 were included

TABLE 1
A DIMENSION OF MANAGERIAL EFFECTIVENESS WITH BEHAVIOUR
EXAMPLES IN RANDOM ORDER

ORGANIZATIONAL AND ENVIRONMENTAL AWARENESS AND SENSITIVITY

Having and using knowledge of changing situations and pressures inside and outside the organization to identify potential problems and opportunities; ability to perceive the impact and implications of decisions on other components of the organization.

1. Seeks understanding of the company's philosophy, its objectives, policy and procedures, the Liquor Act and the business environment.
 2. Understands the organizational hierarchy and the relative importance of departments and persons.
 3. Pretends ignorance when approached by H.O. and instead pushes the responsibility on to some other body.
 4. Awareness of the impact of environmental changes on the company, e.g. (1) the influence of decisions by Regional Councils (Black) on the labour force; (2) changes in the liquor market and other sectors of the economy and the influence thereof on sales figures, etc.
 5. Feeds information from the environment back to the organization, e.g. client who is experiencing financial difficulties.
 6. Does not realize the importance of giving information relevant to a national company activity or problem by just shying away from it.
 7. Is sensitive to factors outside own department which would be beneficial to the company as a whole in terms of savings, efficiencies and profits, e.g. Operations Department re. the use of semi-sweet in the blending of Autumn Harvest, versus semi-concentrate.
 8. Aims at broader objectives than those of his department; willingness to cooperate with other departments.
 9. Refrains from unjustified criticism against departments.
 10. Identifies business opportunities by being alert and well informed.
 11. Uses the right channels of communication.
-

in the 19 scales, with each scale comprising 5 behaviour examples.

Validity and reliability

Determining reliability of BARS is not possible without a comprehensive but separate study. In the present study maximum reliability was ensured by careful construction of the Scales and orientation of raters.

Content validity was built in by emphasizing representativeness of critical incidents. According to senior managers who gave inputs during the construction of the scales, the scales did indeed represent the content of key middle management positions.

Predictive (or concurrent) validity could not, within the objectives of this study, be developed empirically. It will, however, be reflected by validity coefficients obtained through practical use of the BARS.

DISCUSSION

As stated earlier there were two main objectives for developing BARS, namely the validation of an Assessment Centre for middle managers and providing a measurement of on-the-job performance in behavioural terms.

Regarding the first objective, the BARS have been used as an independent criterion in the validation study of an assessment centre (Spangenberg, et. al., 1988). The results of this study seem to confirm the value of BARS as a performance criterion.

Firstly mean scores and standard deviations for the 19 scales (N = 110) were within accepted limits. Mean scores varied from 4.02 to 3.42 (median = 3.62) and standard deviation from .74 to .55 (median = .63).

Feedback from line managers who rated the performance of their subordinates on these scales

indicated that scales were easily understood. Furthermore, behavioural examples enabled raters to use the entire scale. The aforementioned factors brought about a common understanding and standardised application of the scales.

As far as the second, longer term objective is

concerned, the validity study suggests that the BARS can now be made available for wider application. As initially intended, it can be used both for assessing the impact of development activities on the skill levels of assessment centre participants and as a diagnostic tool in identifying performance deficiencies.

TABLE 2
MEANS AND STANDARD DEVIATIONS OF BEHAVIOUR EXAMPLES
FOR THE RATING SCALE OF JUDGEMENT

Judgement		
BEHAVIOUR EXAMPLE	MEAN	STANDARD DEVIATION
2	4,41H	0,42
1	4,12	0,60
7	4,11	0,59
5	4,07MH	0,63
13	3,88	0,79
10	3,49	0,96
9	3,46	0,68
14	3,17M	0,78
16	3,01	0,91
15	2,29	0,58
4	2,19	0,65
12	1,93ML	0,52
11	1,92	0,59
8	1,88	0,63
6	1,67	0,54
3	1,53L	0,48

REFERENCES

- Borman, W.C. & Abrahams, M. (1976). *Developing behaviourally-based performance scales for navy recruiters*. Navy Personnel Research and Development Centre.
- Kotter, J.P. (1982) *The general managers*. New York: The Free Press.
- Langford, V. (1980). Managerial effectiveness – criteria and measurement. *Management Bibliographies and Reviews*, 6, 93-111.
- Luthans, F., & Lockwood, D.L. (1984). Toward an observation system for measuring leader behaviour in natural settings. In J.G. Hunt, D. Hosking, C. Schriesheim & R. Stewart (Eds). *Leaders and managers: International perspectives on managerial behaviour and leadership*. New York: Pergamon Press.
- McCloskey, J.L. (1979). *The development of behaviourally-based scales for the appraisal of managerial job performance*. Ottawa, Ontario, Canada: Public Service Commission of Canada, Personnel Psychology Centre.
- Mintzberg, H. (1973). *The nature of managerial work*. New York: Harper & Row.
- Slivinski, L.W., Grant, K.W., Bourgeois, R.P., & Pederson, L.D. (1977). *Development and application of a first level management assessment centre*. Ottawa, Ontario, Canada: Public Service Commission of Canada, Personnel Psychology Centre.
- 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.
- Spangenberg, H.H. & Esterhuyse, J.J. (1985). *Construction of a Middle Management Assessment Centre (Phase 1)*. Stellenbosch: Stellenbosch Farmer's Wineries.
- Spangenberg, H.H., Esterhuyse, J.J., Visser, J.H., Briedenhann, J.E., & Calitz, C.J. (1988). *Validation of an assessment centre against BARS – an experience with performance related criteria*. Manuscript submitted for publication.