HALO, CENTRAL TENDENCY, AND LENIENCY IN PERFORMANCE APPRAISAL: A COMPARISON BETWEEN A GRAPHIC RATING SCALE AND A BEHAVIOURALLY BASED MEASURE

X.C. BIRKENBACH*

DEPARTMENT OF INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY
UNIVERSITY OF PORT ELIZABETH

OPSOMMING

Die proses van prestasiebeoordeling speel 'n belangrike rol in die ontwikkeling van werknemers asook om administratiewe besluite te maak oor personeel. Die betroubare en akkurate evaluering van werkprestasie kan egter belemmer word deur beoordelingsfoute soos die stralekranseffek, toegeeflikheid, en sentrale neiging. Omrede die alombekende grafiese beoordelingskaal veronderstel is om baie vatbaar te wees vir beoordelingsfoute is daar die afgelope paar jaar aandag geskenk aan die ontwikkeling van gedragsgeoriënteerde beoordelingsmetodes. Dit word aangevoer dat laasgenoemde minder onderworpe is aan beoordelingsfoute. Hierdie studie het die beoordelings van 'n groep werkers op 'n grafiese beoordelingskaal en 'n gedragswaarnemingskaal met mekaar vergelyk. Die resultate kon nie ondersteuning verleen aan die standpunt dat grafiese skale meer vatbaar is vir beoordelings-foute nie.

The importance of performance appraisal as a personnel management function is widely recognised. The greater concern with the systematic evaluation of personnel is easy to understand when the uses that appraisal information can serve are considered. Very often the process of appraisal serves two main functions, namely, administrative and developmental. As far as administrative functions are concerned, performance appraisals are used, for example, to act as criteria in validating selection measures (e.g. tests), assess training needs, evaluate the effectiveness of training, and to make promotion, transfer or termination decisions. In this context the appraisal data are used by management in order to make decisions about people. In this role appraisals are not only used to enhance organizational efficiency but also to serve as a formal record for justifying management actions. The latter issue is important when

* Requests for reprints should be sent to the author

viewed against the background of increasing levels of industrial relations activity with concomitant increases in management actions being challenged by trade unions.

With regard to the developmental functions, formal appraisals assist in providing employees with feedback regarding their job performance as well as providing opportunities to counsel employees on areas of deficient performance. Naturally this is very important to ensure the progressive development of individuals in the organization.

Although the purposes of appraisal cited here provide strong arguments in favour of some form of employee evaluation, in reality a number of factors exist which appear to be limiting the more widespread use of these techniques. A group of factors which appear to be particularly vexing are the so-called rating errors of which halo, central tendency, and leniency have been well-documented in the literature. Such errors are very likely to contaminate ratings of employees with the result that performance scores received by people are rendered inaccurate (Borman 1979).

When succumbing to the halo error, raters apparently assign ratings to individuals by attending to a global impression of each employee rather than distinguishing between performance levels and different dimensions of performance (Borman, 1975). Leniency error refers to the situation where raters generally assign either very high ratings ("easy raters") or very low ratings ("harsh raters") to employees. Central tendency is said to occur when a rater generally refrains from assigning extreme ratings and rather prefers to give average ratings.

Minimising rating errors

At least three possible approaches have been suggested as means for reducing rating errors. Firstly, it has been proposed that errors be corrected statistically. For example, leniency and central tendency could be reduced and even eliminated by adopting the forced distribution ratings according to a normal curve, or by standardizing scores to establish norms of comparability between different raters (Flippo, 1980).

Secondly, training of raters has been put forward as a remedy for the problem. However, it seems as if the empirical evidence for the effectiveness of training in combating rating errors has been mixed. For example, although some research has shown positive effects after training (e.g. Borman, 1975), these effects have not always been upheld on follow-up (e.g. Bernadin, 1978).

It does seem to be certain, however, that merely having knowledge of rating errors (for example, as acquired through a lecture) is not sufficient to change rater behaviour to reduce rating errors. According to Latham and Wexley (1981) the most appropriate techniques in training raters to minimise rating errors are experiential exercises such as modelling, role playing and group discussion.

Finally, combating rating errors has been approached by developing different formats of appraisal. This could probably be seen as the traditional way of approaching the problem as is evidenced by the numerous methods of appraisal which have been developed over the years. The most commonly used technique is the graphic rating scale. Typically this format requires that the rater allocate a numerical value on a scale in terms of his judgement of the performance of an individual on such dimensions as quality of work, quantity of work, leadership or attitude towards the job. Although this format is used very widely, it is the most susceptible to errors in rating. It is particularly because of the proneness of graphic rating scales to rating errors that other formats have been developed. One of the more significant developments in recent years is the appraisal approaches which are based on critical incidents of job performance and described as behaviourally based instruments. Two examples from this category are behaviourally anchored rating scales (Smith and Kendall, 1963) and behavioural observation scales (Latham and Wexley, 1981).

These measures are usually developed through a thorough job analysis in which critical incidents of job performance are gathered from line managers. These incidents are categorised into suitable performance dimensions and form the basis of the appraisal format which will allow supervisors to rate the performances of their subordinates on a suitable scale.

A major advantage of these measures is that rating errors are claimed to be reduced. This advantage seems to occur because levels of performance are defined better, performance dimensions are specified better, and raters are likely to be more cooperative and attentive to the rating task (Borman and Dunnette, 1975). However, research which has investigated the robustness of behaviourally based measures against rating errors has not always supported this contention. In fact Borman and Dunnette (1975) have seriously questioned the usefulness of behavioural measures given the time and effort needed to develop them.

This brief summary of a few relevant issues surrounding appraisal of people at work suggests that although performance appraisal is of considerable importance to both management and employees in an organization, the prevalence of rating errors could significantly reduce the utility of the practice and negatively influence the decisions based on it.

Although a number of approaches have been proposed to reduce the effects of rating errors, this study considered the comparison of results when two different formats were used on the same sample of people when rated by the same group of raters. The first format consisted of a behaviourally based format, namely, a behavioural observation scale (BOS) which was developed along the lines as suggested by Latham and Wexley (1981). The second format was a graphic rating scale which used the same dimension titles as the BOS. These instruments were developed with a view to evaluate the performance of a group of charge hands in a manufacturing organization.

METHOD

Developing of the BOS

The first stage in the development of this format involved the gathering of critical incidents of effective and ineffective work behaviours on the part of the charge hands in the organization. These were obtained by means of interviews conducted by the writer with 17 line supervisors. Approximately 150 critical incidents were obtained in this manner. However, after having grouped similar incidents into single behavioural items and after having eliminated those incidents which were common across interviewees, the total number of incidents were reduced to 51. Furthermore, it was considered that not all of these critical incidents would be equally important in producing effective or ineffective performance and that only those incidents which were really important should be included in the final appraisal format. Following the suggestion of Rossinger et al. (1982), importance was defined in terms of two dimensions, namely, the criticality of the given behaviour, and the frequency of occurrence of the behaviour. In other words, those behaviours which were regarded as being very critical to job success and which also occurred very frequently would be regarded as being more important than behaviours low in criticality and frequency. Importance was determined by presenting a list containing all the critical incidents to 12 of the supervisors who were originally involved in generating the incidents.

These persons were requested to rate each behavioural statement describing a critical incident on a 5-point scale ranging from low criticality to high criticality as well as on a 5-point scale ranging from low frequency to high frequency. The mean ratings on these two dimensions were then computed. On a purely arbitrary basis it was decided to eliminate all the

items that received mean ratings of less than 3 on any of the two dimensions. This exercise resulted in the elimination of 11 incidents.

The remaining number of items were content-analysed in order to group critical incidents which were conceptually similar under a common criterion heading. The 40 behavioural descriptions were categorised to form F criterion dimensions which constituted the final BOS format. The dimensions were:

- Delegation of work (5 items) e.g. "Workers carry out his instructions without arguments".
- Supervision of subordinates (6 items) e.g. "Sees that workers don't fool around on their job stations".
- Technical work performance (10 items) e.g. "Ensures that machine settings are correct in terms of specifications".
- Planning and organizing of work (8 items) e.g. "Keeps a work schedule and checks completed jobs against listed jobs".
- Problem-solving (6 items) e.g. "Reports problems immediately before the situation gets out of hand".
- Safety (3 items) e.g. "Ensures that work areas and aisles are not obstructed or littered".
- Attending work (2 items) e.g. "Stays on the job and does not loiter with workers".

Each item was matched with a 5-point scale on which supervisors were requested to indicate how frequently they had observed the charge engaging in a particular behaviour. Two verbal anchors "almost always"/"almost never" were used on the scale extremes.

Development of the graphic rating scale

The graphic rating scale was developed by simply using the 7 criterion headings of the BOS as the rating dimensions. Brief verbal descriptions defined the various dimensions. Each dimension was matched with a 10-point scale with the verbal anchors of "low ability", "average ability" and "high ability".

Administration of rating formats

Nine of the supervisors who had been involved in the development of the rating formats were asked to rate the performance of the charge hands under their control. The total sample of ratees was 30. The supervisors were first of all requested to rate their charge hands on the graphic rating scale. One week later, after the graphic rating scales had been collected, the same raters were requested to evaluate the performance of the same charge hands using the BOS. The purpose behind these separate rating sessions was to minimise the possibility of one rating event influencing the other.

RESULTS

Relationship between the formats

Prior to the analysis of the rating errors associated with the respective formats, the performance dimensions and summated totals of each format were correlated to see how the two measures were related. These results are given in Table 1.

TABLE 1

CORRELATION BETWEEN DIMENSIONS OF THE GRAPHIC RATING SCALE AND THE BOS

DIMENSION	
Delegation	0,66**
Problemsolving	0,34
Supervision	0,47**
Technical	0,75**
Planning and Organizing	0,42*
Safety	0,42*
Attendance	0,59**
Total score	0,65**

* $p < 0,05$

** $p<0,0l$

An inspection of the coefficients in Table 1 reveals that the degree of association between the graphic rating scale and the BOS ranged from low to moderate. Only one particularly high correlation was obtained, namely, between the two Technical dimensions ($r = 0,75$). This comparatively high correlation is perhaps plausible because technical work behaviour is much more visible than many other forms of work activity. Therefore, it is likely

that technical job performance could be rated relatively objectively and consistently even when using different formats.

In summary it must be deduced that the moderate correlations which were generally obtained show that the two formats did not share a great deal of common variance. Also, because the true performance levels (in psychometric terms) of the employees were not known, it is, of course, impossible to say which of the two formats produced the more accurate results. Indeed, this is always a problem when no independent, external criterion of job proficiency is available against which each of the measures could be judged.

Assessment of halo

In the present study halo was operationally defined in terms of the sizes of the inter-correlations between the dimensions of each format - the greater the sizes of the correlation coefficient, the greater the halo was assumed to be.

The inter-correlation matrix of the performance dimensions for the graphic rating scale and the BOS are given in Tables 2 and 3 respectively.

Firstly, with reference to the figures in Table 2, it will be noted that the correlations between the dimensions vary between 0,22 and 0,79. The mean of the correlations is 0,51. Generally, the sizes of the correlations can be described as moderate. In fact only two correlations can be regarded as being relatively high, that is, over 0,70.

Then, considering the comparable data for the BOS in Table 3, it is immediately clear that the sizes of the correlations are now somewhat higher than for the graphic rating scale. Whereas only two of the graphic rating scale correlations were greater than 0,70, eleven of the BOS inter correlations exceed 0,70. The mean of the correlation coefficients for the respective formats is, however, not statistically significant (t = 0,84).

TABLE 2

INTER-CORRELATIONS BETWEEN PERFORMANCE DIMENSIONS FOR THE GRAPHIC RATING SCALE FORMAT

	1	2	3	4	5	6
1. Delegation	-					
2. Problemsolving	0,27	-				
3. Supervision	0,79**	0,22	-			
4. Technical	0,47**	0,73**	0,43*	-		
5. Planning and Organizing	0,56**	0,59**	0,54**	0,56**	-	
6. Safety	0,40*	0,57**	0,30	0,46*	0,50**	-
7. Attendance	0,60**	0,40*	0,47**	0,46*	0,67**	0,52**

* $p < 0,05$
** $p < 0,01$

Although true halo cannot be determined on the basis of the present data, it is clear that the graphic rating format did not produce more halo (as defined in this study) than the BOS format. Indeed, the results would seem to suggest that the BOS could be more prone to halo than the graphic scale.

TABLE 3

INTER-CORRELATIONS BETWEEN PERFORMANCE DIMENSIONS FOR THE GRAPHIC RATING SCALE FORMAT

	1	2	3	4	5	6
1. Delegation	-					
2. Problemsolving	0,47**	-				
3. Supervision	0,70**	0,77**	-			
4. Technical	0,72**	0,74**	0,77**	-		
5. Planning and Organizing	0,61**	0,71**	0,63**	0,82**	-	
6. Safety	0,22	0,70**	0,66**	0,62**	0,64**	-
7. Attendance	0,57**	0,74**	0,83**	0,69**	0,61**	0,72**

* $p = < 0,05$
** $p = < 0,01$

Assessment of leniency

Leniency in the present study was assessed by investigating the degree of skewness of the distributions for each performance dimension of the respective rating formats. The skewness measures are presented in Table 4.

A skewness score of zero means that the spread of ratings is in the form of a normal distribution. A positive score implies that the distribution is positively skewed, that is, symptomatic of "harsh" ratings. A negative score implies a negatively skewed distribution, that is, "easy" ratings. Naturally, the larger the sizes of the skewness' scores, the greater the degree of skewness.

TABLE 4

SKEWNESS SCORES FOR THE PERFORMANCE DIMENSIONS ON THE GRAPHIC RATING SCALE AND THE BOS

DIMENSION	GRAPHIC RATING SCALE	BOS
Delegation	- 0,38	- 0,23
Problemsolving	0,15	0,32
Supervision	- 0,21	- 0,50
Technical	- 0,47	- 0,48
Planning and Organization	- 0,03	- 0,06
Safety	- 0,59	0,69
Attendance	- 0,57	- 0,24
Total score	- 0,73	- 0,35

A number of interesting observations can be made from the information in Table 4. First of all, it is apparent that the scores are generally negative. This implies that both measures generally gave rise to "easy" ratings. In only one case, namely for Problemsolving did both measures produce positively-skewed distributions. The fact that both formats produced positive scores does seem to imply that the raters were generally relatively critical of their subordinates' abilities to solve work-related problems.

A second striking observation which can be made from the data in Table 4 is the scores for the performance dimension of safety. In this case a reversal of scores were recorded; that is, while the graphic rating scale produced an "easy" distribution, the BOS produced a comparatively "harsh" distribution. Therefore, it can be deduced that the BOS format possibly allowed the rater to more critically evaluate his subordinates' safety behaviours as opposed to the more global assessment of the graphic rating scale.

Apart from the single anomaly described, it would appear that the two formats generally produced similar trends of skewness. However, when looking at the sizes of the differences between each pair of scores it does seem that the graphic rating scale produced "easier" distributions on the average compared to the BOS. In fact, the graphic rating scale has higher negative scores for all the performance dimensions excluding Supervision, Technical, and Planning and organization. (The differences for the latter two are virtually zero, however).

Assessment of central tendency

Central tendency was assessed by a study of the kurtosis of the distributions produced by each performance dimension for the respective formats. A negative kurtosis score is indicative of a platykurtic curve, that is, a flat curve which has ratings distributed across the range of possible scores. Positive kurtosis on the other hand is referred to as representing a leptokurtic distribution, that is, a tall curve with ratings distributed only within a narrow range of scores. This type of distribution is suggestive of central tendency error in performance appraisal.

The kurtosis scores for the variables in the study are given in Table 5.

TABEL 5

KURTOSIS SCORES FOR THE PERFORMANCE DIMENSIONS ON THE GRAPHIC RATING SCALE AND THE BOS

DIMENSION	GRAPHIC RATING SCALE	BOS
Delegation	- 0,89	- 0,96
Problemsolving	- 1,04	- 0,67
Supervision	- 1,00	- 0,79
Technical	- 0,79	- 0,98
Planning and Organization	- 0,94	- 1,07
Safety	- 0,84	- 0,16
Attendance	- 0,49	- 0,89
Total	0,55	- 0,92

It is evident from the data presented in Table 5 that all the kurtosis scores, barring the total score for the graphic rating scale, showed negative kurtosis. Clearly, therefore, raters tended to produce flat curves on both the formats. Although the sizes of the respective

kurtosis scores vary between the two rating instruments, the similar trends are, nevertheless, striking.

The only anomaly is the kurtosis score recorded for the total summated score on the graphic rating scale. Compared to the total summated score for the BOS, the former shows a relatively high degree of central tendency. However, the fact that this high degree of central tendency occurred on the summated score of the graphic rating scale does seem to suggest that raters did not consciously allocate ratings to a central point on the rating continuum. Thus, it may be possible that this result is spurious.

To summarise these results, it is evident that both formats almost exclusively produced flat curves as opposed to tall curves. Thus, neither format gave rise to central tendency.

CONCLUSION

Although the literature on behaviourally based performance measures has suggested that these formats are less susceptible to rating errors than the widely used graphic rating scale, the results of this study have not lent support to this contention. In fact, the distributions of the performance dimensions for the respective measures were remarkably similar. For example, both methods exclusively produced "flat" curves which indicated that neither instrument was prone to the central tendency error. Also, as far as skewness was concerned, the distributions obtained generally followed similar trends which can be interpreted as evidence for similarity in leniency - both formats on the average produced "easy" distributions as opposed to "harsh" distributions. The area where the greatest discrepancies between the two formats occurred was on the measures of halo. The BOS method almost consistently produced higher inter-correlations between dimensions than did the graphic rating scale. Intuitively this could be interpreted as evidence that the BOS was more susceptible to halo than the graphic rating scale. However, this interpretation need not necessarily be correct and should be qualified. An important matter to consider when these differences are evaluated is that each format produced an *observed* score. The *true* performance scores for each ratee were of course, not known. Therefore, high inter-correlations between dimensions for any given format need not necessarily be an indication of halo - in fact, this may in reality be a reflection of the true state of affairs. Thus, on the basis of the present data it cannot equivocally be stated that the BOS format gave rise to more halo error than the graphic rating scale. However, the nature of the correlations does allow one to deduce that the graphic rating scale did not

produce *greater* halo than the BOS as has been intimated in the literature on appraisal (e.g. Vance et al, 1978).

Against this background the pertinent question originally addressed by Borman and Dunnette (1975) again arises, namely, how to evaluate the utility of behaviourally based appraisal methods given the time and effort required to develop them. As far as this study is concerned, the question has to be approached with at least two factors in mind. First of all, the study merely investigated the effects of format on rating distributions. According to Latham and Wexley (1981) a more important variable which needs to be addressed in order to reduce rating errors is the training of raters. These authors have concluded that when raters have not received any training in performance evaluation, it does not matter which type of rating format is used as both are likely to produce similar results.

However, "after supervisors are trained to improve their objectivity in observing and recording employee behaviour, trait scales (graphic rating scales) are clearly inferior to other scales" (Latham and Wexley, 1981, p.38). The implication of this statement for the present study is that dissimilar patterns of ratings could perhaps have emerged for the two formats if raters had been trained in evaluation of performance.

A second factor which has to be borne in mind when evaluating the utility of the BOS is the use to which the ratings will be put. Perhaps the most important use of appraisal data is to develop individuals in the organization. Now, although the results of the study showed that there were no major differences in the distributions and correlations of the respective ratings, the actual use to which the BOS ratings could be put in facilitating the development of ratees should be clear. In other words, whereas the graphic rating scale will only allow a superior to tell his subordinate that his performance was below average/above average on a specific dimension, the BOS makes it possible for the superior to clearly pin-point those important behaviours which were manifested effectively or ineffectively and which were responsible for the person receiving a given rating. Thus, in contrast to the graphic rating scale, the BOS format allows specific feedback to be given to improve weak areas or to reinforce strong areas of work performance.

What this really amounts to is that while these results have suggested little *statistical* differences between the two formats, there appears to be a significant *behavioural* advantage attached to the use of behaviourally based performance appraisal formats. This is possibly sufficient justification for adopting a behaviourally based method of appraisal such as the BOS.

ABSTRACT

The process of performance appraisal can serve important employee development as well as organizational administrative functions. However, the reliable and accurate assessment of performance could be hampered by rating errors such as halo, leniency, and central tendency. Because the traditional approach to appraisal by means of graphic rating scales is considered to be susceptible to these errors, behaviourally based measures have been developed which have the claimed advantage of being relatively resistant to rating errors. This study compared the ratings given to a group of employees on a graphic rating scale and a behavioural observation scale. The results did not support the superiority of the BOS in resisting rating errors.

REFERENCES

Bernadin, H.J. Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 1978, *63*, 301-308.

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

Borman, W.C. & Dunnette, M.D. Behaviour-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, 1975, *60*, 561-565.

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

Flippo, E.B. *Personnel Management*. Tokyo: McGraw-Hill Ltd., 1980.

Latham, G.P. & Wexley, K.N. *Increasing productivity through performance appraisal*. Reading: Addison Wesley, 1981.

Rossinger, G., Myers, L.B., Levy, G.W., Loar, M., Mohrman, S.A., & Stock, J.R. Development of a behaviourally based performance appraisal system. *Personnel Psychology*, 1982, *35*, 75-88.

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.

Vance, R.J., Kuhnert, K.W. & Farr, J.L. Interview judgements: Using external criteria to compare behavioural and graphic scale ratings. *Organizational Behaviour and Human Performance,* 1978, *22*, 279-294.