DOCUMENT RESUME

ED 141 507 CE 011 068

AUTHOR Schwind, Hermann F.
TITLE New Ways to Evaluate Teaching and Training Effectiveness.
PUB DATE Apr 77
NOTE 31p.; Paper presented at the Adult Education Research Conference (Minneapolis, Minnesota, April 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Behavioral Objectives; Behavior Patterns; *Behavior Rating Scales; Behavior Standards; Evaluation Criteria; *Evaluation Methods; Literature Reviews; *Measurement Instruments; Performance Specifications; Personnel Evaluation; *Task Performance; Test Reliability

ABSTRACT

This paper discusses the advantages and disadvantages of commonly used measures of job effectiveness, concentrating on a recent development in the field, the Behaviorally Anchored Rating Scale (BARS); and proposes a new approach, the Behavior Description Index (BDI), which the author contends reduces or avoids most of the shortcomings of other methods. After discussing the advantages and distinguishing features of BARS, the author refers to the main problems of currently used instruments that have been cited in the literature: Low inter-rater reliability, central tendency (inclination of rater to avoid extreme ratings), halo effect (tendency to assign the same rating to each factor being rated), and leniency effect (tendency of supervisors to overrate subordinates). Two shortcomings not dealt with in the literature reviewed are also presented: Waste of valuable information and multidimensionality. The paper then examines the characteristics of the BDI and claims advantages of the new scale over other scales (for example, that the BDI uses behavioral criteria; uses a larger sample of the total job behavior domain than BARS; has less leniency, halo, and central tendency effects; and probably has higher inter-rater reliability). Implications of the use of the new instrument in performance and training evaluation are discussed. References and examples of statements from the BARS are appended. (LMS)

NEW WAYS TO EVALUATE

TEACHING AND TRAINING EFFECTIVENESS

by

Hermann F. Schwind
Assistant Professor
Faculty of Commerce
SAINT MARY'S UNIVERSITY
Halifax, Nova Scotia
B3H 3C3

April, 1977

ABSTRACT

The main problems and disadvantages of currently used performance appraisal and training evaluation methods are discussed: leniency, central tendency, and halo effects, low interrater reliability, and improper criteria (low validity). A new approach to performance and training evaluation is the use of critical incidents in Behaviourally Anchored Rating Scales (BARS). It is claimed that BARS have more positive characteristics than other scales; e.g. summated rating scales, graphic rating scales, etc. However, recent research has shown that under certain conditions BARS do not demonstrate superior qualities to other scales. An improved use of critical incidents is described which has all the advantages BARS possess, but avoids the disadvantages. The new instrument is the Behaviour Description Index (BDI) which uses behavioural criteria, utilizes a larger sample of the total job behaviour domain than BARS, has less leniency, halo, and central tendency effects, and probably higher interrater reliability. Implications of the use of the new instrument in performance and training evaluation are discussed.

With the economic slump during the last few years, and resulting budget cuts for educational institutions and industrial training programmes, came a new interest in the assessment of the quality of instructors and the outcome of training and development programmes in government and industry. While research concerned with evaluation of teaching effectiveness has been a going concern for decades, the evaluation of training programmes "is much in the same category Mark Twain placed the weather," to use McGehee and Thayer's words (1961, p. 256). Everybody talks about it, but little is done. There are several possible reasons for this, such as

1. Difficulty in doing controlled studies in organizations;
2. Costs of training evaluations;
3. Fear of discovering unwelcome facts on training outcomes;
4. Unwillingness to accept new approaches, a "don't rock the boat" attitude.[1]

A seemingly important reason is not mentioned here: the lack of a valid and reliable evaluation instrument, a fact which seems to be true also for the evaluation of teaching effectiveness.

A look at currently used evaluation forms to measure "teaching ability" illustrates the problem. As Harari and Zedeck (1973) put it: "Current...evaluation forms are often ambiguous, verbose, disorganized and arbitrarily

developed. They consist of global behavioral measures and vague trait descriptions. As a result, the forms tend to be unreliable and very susceptible to response biases."

The situation is worse in the area of training evaluation. The majority of business organizations seem to use simple reaction measurements (How did you like the program?) to assess the outcome of their training programs (Catalano and Kirkpatrick, 1968). It should be obvious that the fact that a participant likes a certain programme says nothing about its effectiveness.

An instructor's effectiveness and the outcome of an educational or training programme are certainly closely related, but we are dealing with two different concepts. When we measure teaching effectiveness we assess an instructor's teaching behaviour while the evaluation of a programme assesses the degree of a participant's behaviour change. The assessment of the first is based on judgements made either by students, colleagues or superiors, while the assessment of the second is based on observations of job related behavour by superiors or neutral observers.

This paper discusses the advantages and disadvantages of commonly used effectiveness measures and will concentrate on the most recent development in this field, the

Behaviourally Anchored Rating Scale (BARS). A new approach, the Behaviour Description Index (BDI), will be proposed which will reduce or avoid most of the shortcomings of past and current methods.

The Critical Incident Method

In his search for a useful tool to measure work performance, Flanagan (1954) developed a technique which he called the Critical Incident Method. He defined the critical requirements of a job as those behaviours which are crucial in making a difference between doing the job effectively and doing it ineffectively. Critical incidents, as the term implies, are simply reports by qualified observers of the things people did that were especially effective or ineffective in accomplishing parts of their jobs. Such incidents are actual behavioural accounts, recorded as stories or anecdotes and obtained from managers, job incumbents, and others close to the job being studied. In its simplest form the Critical Incident Method consists of listings of critical incidents which are then compared to exhibited behaviour of employees to be rated. The underlying objective of this procedure is to obtain a job-specific scale of behaviour effectiveness. A modified version of this approach is the Behaviourally Anchored Rating Scale(s) (BARS).

Development of Behaviourally Anchored Rating Scales

The first step in developing BARS is to ask a sample of job incumbents and/or their immediate supervisors and -- where applicable -- subordinates to write short descriptions

of an incumbent's job behaviour they have either observed or heard about. Each behavioural description is characterized by the specification of a single job situation and the behaviour in response to that situation. To be critical, an incident must occur in a job situation where the purpose or intent of the behaviour is fairly clear and where its consequences are definite enough to leave little doubt concerning its effect on job performance (Flanagan, 1954). Although the latter definition suggests description of extreme behaviour, a critical incident need not refer exclusively to the extremes of performance.

After they are gathered, the pool of behavioural incidents are usually edited to conform to an expected behaviour format, that is, each description of an incumbent's job behaviour is prefaced with the phrase "Could be expected...." The intent of the phrase is to allow the rater to generalize from what he has seen the ratee do in a situation to what he would expect the ratee to do in a particular situation, regardless of the opportunity to actually observe the ratee.

After the editing and rephrasing, the incidents are categorized. Usually the researcher reads, sorts, and then labels groups of incidents in terms of similarity, or the researcher first qualitatively identifies a set of dimensions and then sorts statements according to their

similarity in meaning to the a priori defined dimensions. To avoid criterion contamination through personal biases, an incident reallocation or "retranslation" procedure is used (Smith and Kendall, 1963). In the retranslation process experts (job-knowledgeable employees) are provided with a list of job dimension definitions and asked to assign the incidents to the behavioural dimensions they feel they describe. Criteria of retention are included in the procedure for determining the extent to which a particular incident is part of a dimension. Those incidents meeting the criteria are retained for subsequent use.

Following the retranslation procedure the incidents are rated on a Likert-type scale (usually from 1 to 7, or 1 to 9) as to the degree of effectiveness they characterize in each job dimension.

The mean rating for an incident determines its scale value, while the Standard Deviation (SD) of the mean rating is viewed as an index of ambiguity. The procedure requires the ambiguous incidents (i.e., incidents with an SD in excess of some minimum value) be excluded from the scales. The retained incidents are then ordered within the performance dimensions they anchor in terms of their mean scale values. The usual arrangement of anchors is a vertical graphic scale, consisting of a vertical line, market in equal-appearing intervals, with incidents arranged along

its length according to their mean value. Each scale is headed by a dimension definition and usually omits the scale value for each incident of behaviour. Exhibit 1 is an example of a scale developed by Das (1975), for instructors of a school of business administration.

In using the scale, the rater is instructed first to read the dimension definition for a scale. Then, he is asked to read each incident starting at the bottom and reading toward the top until he reads an incident that exceeds the ratee's "typical" best job behaviour. He then returns to the highest "typical incident" and checks it as indicative of the ratee's job performance within that dimension of the job. The value of the incident checked by the rater determines the performance score on that dimension.

It is possible to score the scales in at least two ways. First, if a summation is desired, scores can be summed across dimensions for a ratee. Second, if a performance profile is desired, the score on each scale can be reported. Typically, the latter format is used.

Advantages and Distinguished Features

Zedeck, et al. (1973, p. 1) list the following points as the advantages and distinguishing features of BARS procedures:

1. Each scale employs job-related behavioural incidents as anchors or reference points;

2. Groups with work experiences similar to those who eventually use (or are subjected to) the scales participate... (in their development) ...;

3. The terminology commonly used in the... (job to be rated)...is retained in the anchors;

4. A (reallocation) procedure is used...to reduce the ambiguity of the scales...;

5. Conceptually independent scales with high scale reliability are obtained; and

6. In actual use, ratings...(can be)... documented with specific incidents....

Smith and Kendall (1963) suggest that their retranslation procedure will lead to less leniency and central tendency errors. Cummings and Schwab (1973) point out that the use of critical incidents may prove to be useful in providing feedback to appraisers, since their specificity can serve as a concrete example of areas where job behaviours could be improved.

Inter-rater Reliability

Campbell, et al. (1973) and Zedeck and Baker (1972) assessed the inter-rater reliability of specific BARS instruments. In both studies, inter-rater reliability was low to moderate (i.e., r's ranged from .24 to .55). The results suggest that acceptable levels of inter-rater reliability have not yet been obtained using BARS procedures.

In both studies, however, it may have been that the tests of inter-rater reliability were deficient, that coefficients were computed between levels of supervision, rather than within. As Campbell, et al. (1973) suggest, perhaps the different levels of supervision had different opportunities to observe ratee behaviour or differed in perceiving the utility of specific behaviours for meeting job requirements. A similar view was expressed by Borman (1974). In his opinion, one should not expect high reliability under such circumstances, since raters at different levels may have different ratee performance dimensions which they feel are relevant. Borman tested his hypothesis by having secretaries and academic instructors each independently develop their own critical incidents for the job of secretary. Both groups identified different performance dimensions. BARS developed from the critical incidents were then used by the two groups to rate the secretaries on all performance dimensions. For each rater group, Borman found that inter-rater reliability was higher for the performance dimensions the group had identified than for the dimensions identified by the other group. This explanation of the low to moderate inter-rater reliability of BARS is plausible. However, there is a second possible explanation. It will be discussed under "Shortcomings of BARS."

Central Tendency

No studies on BARS so far seem to have paid much attention to the problem of central tendency. It describes the inclination of raters to avoid extreme ratings on a scale, e.g. "Outstanding", or "Very Poor." BARS should be less prone to this problem because of their behaviour specificity. Since often independent behaviour samples are utilized, a rater has to make a choice; i.e. he is able to avoid extreme ratings. (See chapter on "Shortcomings of BARS")

Research so far has shown that BARS seem to have a slight advantage over other methods of performance evaluation. However, the question has to be asked whether these findings illustrate a genuine lack of superiority or whether the methodology used in the comparison studies is less than adequate. As mentioned above, Schwab, et al. (1975) criticize the use of only two instruments for comparison purposes. Another problem may lie in the standard deviation criterion used to select the critical incidents. In most cases the criterion was set so that it indicated a substantial amount of disagreement among judges as to the level of effectiveness the behaviour described, typically 1.50, 1.75 or even 2.0 standard deviations. This may suggest that the critical incidents selected for the instrument were not the unambiguous behaviour samples the creators had hoped for. A third questionable area may be calculation

and hence provide somewhat more information about the actual nature of the group being evaluated.

Halo Effect

The halo effect appears in evaluation when the evaluation tends to assign the same rating or level to each factor being rated (Glueck, 1974).

Burnaska and Hollman (1974) found that BARS resulted in less halo than a numerically anchored and adjective scale. However, they point out that all three scales had excessive levels of halo. In a study comparing BARS and a numerically anchored scale, Campbell, et al. (1973) found that the former scale format showed less halo than the latter. Similar results were reported by Groner (1974), Borman & Dunnette (1975), and Keaveny and McGaun (1975). On the other hand, Borman and Vallon (1974) found no differences in halo effect between BARS and other non-behavioural scales. Similar to the critique on the approach to measure leniency, Schwab, et al. (1975) argue against the use of only two instruments to study the halo effect:

> "If one begins with the reasonable assumption that performance on various dimensions is inter-related, then comparison of the intercorrelations generated by just two instruments provides little basis for deciding the actual or true inter-relations in the group appraised." (p. 560)

Supervisors tend to overrate their subordinates; i.e. the tendency is to be lenient rather than strict. The result of this "error" is that the average performance rating is not at the midpoint of a scale but on the positive side of it.

Smith and Kendall (1963) suggest that BARS should be less susceptible to leniency effects because of the unambiguous dimensions and anchors developed by the procedure. However, research results are equivocal on this issue. Campbell, et al. (1973) found that the mean ratings on their instrument were, on average, closer to the midpoints of the scales than those of a summated rating scale. On the other hand, Borman and Vallon (1974) found that a group of employees had significantly higher ratings on a behaviourally anchored scale. Campbell, et al. interpreted their results to mean that BARS demonstrated less leniency while Borman and Vallon concluded that BARS showed greater leniency error. Schwab, et al. (1975) point out that it may be risky to make inferences about relative leniency effects using only two instruments in a comparative study because "it is not possible to determine what the true average rating should have been" (p. 559). For this reason, Schwab, et al. (1975) recommend using more than two sets of measures in the evaluation process. A greater number of measures would allow more comparisons

raters consisting of employees from different levels; e.g. superiors and subordinates of ratees. If the above-mentioned problems are corrected, BARS may demonstrate a more significant advantage over other measures.

Shortcomings of Behaviourally Anchored Rating Scales

In the introduction, several references have been cited which suggest that BARS have a number of advantages over traditional performance rating methods. The advantages claimed are: higher job specificity, higher motivation of raters, higher acceptance of ratings by ratees, higher dimension independence, less halo, less leniency, and less central tendency. However, there seem to be at least two shortcomings of the BARS technique which have not been discussed in the previous literature review:

1. waste of valuable information;
2. multidimensionality.

Waste of Valuable Information

After development, critical incidents for a job are put through the validation and retranslation process, and they must fulfil the standard deviation criterion. Usually 20 to 50 critical incidents per job dimension survive. Yet only between 5 and 10, depending on the number of anchoring points of the scale, are utilized, all others are thrown out. Undoubtedly, those items which are not used contain valuable

information about the job dimension to which they were attributed in the retranslation process. The decision to eliminate them is made on the basis of arbitrarily chosen criteria: a convenient mean value to fit the scale points, and the degree of agreement between raters as measured by the standard deviation.

Multidimensionality

A second problem with BARS has to do with the use of independent critical incidents in behaviour dimensions. Behaviour dimensions are important aspects of a total job behaviour domain which, in turn, is composed of all possible relevant job behaviours. "This instructor always uses the blackboard to illustrate a problem" is a sample of the behaviour dimension "instructor in class" of an instructor's job behaviour domain. Wallace and Schwab (1973) found five behaviour dimensions for an instructor in a school of business (see Exhibit 2), while Das (1975) identified 18 dimensions (see Exhibit 3). When critical incidents are generated, the intention is to sample to a significant degree the behaviour domain of a job dimension. The problem is that there are often so many job dimensions that it is impractical to develop scales for each one, since a rater very likely will refuse to evaluate a ratee on 30 or 40 dimensions. For this reason, job dimensions usually are collapsed into a manageable number, e.g. 5 or 10. As a consequence of this approach most scales utilizing

critical incidents use independent behaviour samples, thus forcing the rater to make a difficult choice, opening up the rating procedure to possible biases, like leniency and halo.

To illustrate the problem, an example is taken from Das (1975). He identified 18 job behaviour dimensions of an instructor. A BARS for one of the dimensions is shown in Exhibit 4. A comparison of the seven behaviour samples reveals that behaviour #1 is conceivably independent of the behaviours #2, 3, and 4. With other words, it is possible that a rater can choose all these behaviours as "typical" and not just one. On the other hand, behaviour samples #1, 5, 6 and 7 are mutually exclusive (dependent). Ideally, a behaviour dimension consists only of mutually exclusive or unidimensional behaviours. Otherwise the rater has to choose between different possible behaviours which leaves the instrument open to response biases and will result in low reliability.

Multidimensionality very likely is also one of the causes of the central tendency effect BARS seems to exhibit, although to a lesser degree than other common scales (Campbell, et al., 1973). Since a rater has independent choices, it is possible for him to avoid extreme ratings. The same characteristic may also be the cause of the halo error.

Characteristics of the New Scale

Instead of the usual utilization of only 5 - 9 behaviour descriptions in a BARS, it is proposed that a larger sample of the total behaviour domain of a job dimension be used (see Exhibit 5). The number of critical incidents could be determined by the total number of critical incidents generated per job dimension. The number of incidents utilized per scale will be limited only by fatigue effects of raters. It is expected that for practical purposes the maximum number of behaviour descriptions will be equal to or less than 20. The utilization of a larger number of critical incidents would overcome or at least reduce one of the major shortcomings of BARS discussed before: the loss of information. It is conceivable that in many instances 20 critical incidents will encompass the total behaviour domain of a job dimension. If it is not the case, then at least a much better sampling can be done. If the total domain were 40, then 20 items represent 50% as compared to 17.5% with a sample of 7.

In order to avoid the disadvantages of graphic rating scales and BARS, a forced choice scoring is suggested. Raters will respond to the question: Does the ratee exhibit the below described behaviour, yes or no? In the case the

rater is not sure, he can respond with a question mark. After the response sheet is completed, all ratings will be converted into points. (For details see chapter "The Rating Procedure for the BDI".) This conversion could be done by a different person than the rater or through a computer program. This approach should have a drastic influence on the halo effect. Since the rater does not know whether he evaluated the ratee high or low on a scale, it is very difficult or impossible for him to transfer a general characteristic from one scale to another. It could be argued that the rater will know how he evaluates a ratee if he responds only positively to critical incidents. However, since positive and negative critical incidents will be randomly mixed in a scale the rater must really concentrate on his responses if he wants to bias the evaluation. He still does not know the actual score (see Rating Procedure for the BDI).

Another possible advantage of the new scale would be a reduced or eliminated leniency effect, largely because of the same reasons described above: mixing of a larger number of positive and negative statements and independent scoring. Only a consciously false response could induce a leniency effect. But no scale is immune against wilful misuse.

A fourth improvement as compared to traditional BARS would be the virtual elimination of the central tendency effect. Since the BDI does not use continuous scales with extreme ratings on either end, the cause of any possible central tendency effect is removed.

There may be a fifth advantage of the BDI over BARS. It has been suggested that the low to moderate inter-rater reliability of BARS may be caused by using raters from different organizational levels (e.g., Campbell, et al., 1973). Nobody so far has pointed to the possibility that the multidimensionality problem may be either a contributing or even a major factor in causing the low reliability. It is quite possible that the BDI will reduce the problem since the rater is not forced to choose one from several possible items, but he may check off as many items as are available. A superior could determine in advance what scores would be acceptable or unacceptable, e.g., out of 60 possible points: (20 items x 3 points)

0 - 30 may mean: urgent training required
(or, if measured after training:
training ineffective)

31 - 40 may mean: training recommended

41 - 50 may mean: refresher course may be useful

51 - 60 may mean: no training required.

In summary, the new BDI scale seems to offer the following advantages:

1. increased information content by improving sampling of the behaviour domain;
2. reduced or no halo effect;
3. reduced or no leniency effect;
4. no central tendency effect;
5. higher inter-rater reliability.

Rating Procedure for the BDI_2

The BDI uses positibe and negative critical incidents in random order. The number of positive and negative statements does not influence the score because of the scoring characteristics. If a rater responds positively (Yes) to a positively worded statement, or negatively (No) to a negatively worded statement, the score will be 3 points. A positive response to a negative statement and vice versa results in 0 points. If the respondent is not sure or cannot decide, the response is a question mark (?) and the score will be 1 point. Again, it will be emphasized that the conversion of the ratings (Yes, No, ?) to point scores (3, 0, 1) is probably not done by the rater, but a second person, or more likely by a computer, especially if the number of ratees is large. It is assumed -- pending empirical investigation -- that the job dimensions are relatively independent. For this reason the point scores will be totalled for each scale separately.

CONCLUSION AND IMPLICATIONS

A new approach to teaching and training evaluation has been discussed. The characteristics of the new instrument -- the Behaviour Description Index -- seem to be superior to conventional performance appraisal, e.g. summated or graphic rating scales and BARS.

If the instrument can be empirically validated -- there is little doubt that it will -- it should prove to be a significant improvement in the evaluation process. Instead of relying on vague trait characteristics which mean different things to different people, very specific behaviours are described, which can easily be observed.

Secondly, a large part of the total behaviour domain can be utilized, enabling raters to pinpoint shortcomings of instructors or training participants, thus making it easier to take corrective actions, e.g. retraining or counselling.

There are other possible uses of the BDI. Much has been written about the vagueness of job descriptions. What could be a better solution to this problem by handing a new job incumbent together with the description of his responsibilities a copy of a BDI of his job? He would find samples of effective and ineffective job behaviour

and would know immediately what is expected of him. Other areas of application could be performance appraisal and determining of training needs. Actually, the BDI could be the basis for a new systems approach in the personnel management area. Future research will show whether this is possible. The first indications are certainly encouraging.

FOOTNOTES

1. For a detailed discussion of these problems, see Schwind, 1975, a and b.

2. This is a similar approach as described in Smith, Kendall and Hulin, 1969.

References

Borman, W.C., "The ratings of individuals in organizations: An alternative approach." Organizational Behavior and Human Performance, 12 (1974) 105 - 124.

Borman, W.C., and Dunnette, M.D., "Behavior based versus traitoriented performance ratings: An empirical study." Journal of Applied Psychology, 60 (1975) 561 - 565.

Borman, W.C., and Vallon, W.R., "A view of what can happen when behavioral expectation scales are developed in one setting and used in another." Journal of Applied Psychology, 59 (1974) 197 - 201.

Burnaska, R.F., and Hollman, T.D., "An empirical comparison of the relative affects of rater response biases or three ratings scale formats." Journal of Applied Psychology, 59 (1974) 307 - 312.

Campbell, J.P., and Dunnette, M.D., and Arvey, R.D., and Hellervik, L.N., "The development and evaluation of behaviorally based ratings scales." Journal of Applied Psychology, 57, (1973) 15 - 22.

Catalano, R.F., and Kirkpatrick, D.L., "Evaluating Training Programs" Training and Development Journal, May 1968.

Cummings, L.L., and Schwab, D.P., Performance in Organizations: Determinants and Appraisal, Glenview, Ill.: Scott, Foresman and Co., 1973.

Das, H., "Behaviorally Anchored Rating Scales in Teaching Evaluation", Unpublished Master Thesis, University of British Columbia, Faculty of Commerce, 1975.

Flanagan, J.C., "The critical incident technique." Psychological Bulletin (1974) 327 - 358.

Glueck, W.F., Personnel, A Diagnostic Approach, Dallas, Texas: Business Publications, 1974.

Groner, D.M., "Reliability and susceptibility to bias of behavioral and graphic rating scales." Unpublished doctoral dissertation, University of Minnesota, 1974.

Harari, D., and Zedeck, S., "Development of Behaviorally Anchored Rating Scales for the Evaluation of Faculty Teaching." Journal of Applied Psychology 58 (1973), 261 - 265.

Keaveny, T.J., and McGaun, A.F., "A comparison of behavioral expectation scales and graphic rating scales." Journal of Applied Psychology, 60 (1975) 695 - 703.

McGehee, W., and Thayer, P.W., "Training in Business and Industry" John Wiley & Sons, Inc., N.Y. 1961

Schwab, D.P., and Heneman, H., and DeCotiis, T.A., "Behaviorally anchored ratings scales: A review of the literature." Academy of Management Proceedings (1975) 222 - 224.

Schwind, H.F., "Thoughts on Training Evaluation." Canadian Training Methods, 8 (1975) #1, 14 - 15.

Smith, P.C., and Kendall, L.M., "Retranslation of expectations: An approach to the construction of unambiguous anchors for ratings scales." Journal of Applied Psychology, 47 (1963) 149 - 155.

Smith, P.C., Kendall, L.M., and Hulin, C.L., "The Measurement of Satisfaction in Work and Retirement." Rand McNally & Co., Chicago, Ill., 1969.

Wallace, M.O., and Schwab, D.P., "The Validation of Teaching-Effectiveness Measure in Two Business Schools" Academy of Management Proceedings, 1973.

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

Zedeck, S., and Imperato, N., and Krausz, M., and Oleno, T., "Development of behaviorally anchored rating scales as a function or organizational level." Journal of Applied Psychology, 59 (1974) 249 - 252.

Exhibit 1

Job Dimensions of Department Manager

1. Supervising sales personnel
2. Handling customer complaints and making adjustments
3. Meeting day-to-day deadlines
4. Merchandise ordering
5. Developing and planning special promotions
6. Assessing sales trends and acting to maintain merchandising position
7. Using company systems and following through on administrative operations
8. Communicating relevant information to associates and to higher management
9. Diagnosing and alleviating special department problems.

from Campbell, J.P., Dunnette, M.D., Arvey, R.D., and Hellervik, L.V., "The Development and Evaluation of Behaviorally Based Rating Scales." Journal of Applied Psychology, Vol. 57, No. 1 (1973) 15-22.

EXHIBIT 2

DIMENSIONS OF TEACHER BEHAVIOUR IDENTIFIED FROM INCIDENTS GIVEN BY STUDENTS

1. Instructor in class
2. Required reading
3. Subject matter
4. Instructor in general
5. Assignments and examinations

EXHIBIT 3

DIMENSIONS (CATEGORIES) OF TEACHER BEHAVIOUR

IDENTIFIED FROM INCIDENTS GIVEN BY STUDENTS

1. Course Outlining and Structuring
2. Administrative Handling
3. Coverage Of Material
4. Teaching Style
5. Teaching Methods
6. Evaluation
7. Interaction Outside Class
8. Flexibility and Responsiveness

DIMENSIONS (CATEGORIES) OF TEACHER BEHAVIOUR

IDENTIFIED FROM INCIDENTS GIVEN BY PROFESSORS

9. Interaction With Colleagues
10. Interaction With Students Outside Class
11. Behaviour In The Class Room
12. Research Activities
13. Handling Administrative Matters

DIMENSIONS (CATEGORIES) OF TEACHER BEHAVIOUR

IDENTIFIED FROM INCIDENTS GIVEN BY ADMINISTRATIVE STAFF

14. Facilitation Of Administrative Work Flow
15. Controlling Expenditures
16. Adherence To Policies
17. Providing Feedback To The Staff
18. Counselling Activities

EXHIBIT 4

BEHAVIOURALLY ANCHORED RATING SCALE

Could be expected to specifically state requirements for the course and use definite and stated criteria for evaluation.

Could be expected to give course outlines and schedules to students.

Could be expected to hold an introductor session in which the students' expectations are ascertained and instructor's course objectives are made clear.

Could be expected to set specific targets for each session.

Could be expected to take quite a few days to tell the students what is the course content of the course.

Could be expected to set class participation as an evaluation criterion without clearly stating his expectations.

Could be expected to announce mid-way through that there would be a final exam, in contradiction to his earlier statement that there would be no final