10-1979

The Operational Effectiveness of the Behavioral Expectations Scale & the Mixed Standard Scale: A Comparative Evaluation

Deborah Boniske
Western Kentucky University

Recommended Citation

Boniske, Deborah, "The Operational Effectiveness of the Behavioral Expectations Scale & the Mixed Standard Scale: A Comparative Evaluation" (1979). *Masters Theses & Specialist Projects.* Paper 2167.
https://digitalcommons.wku.edu/theses/2167

Boniske,

Deborah J.

1979

THE OPERATIONAL EFFECTIVENESS OF THE BEHAVIORAL EXPECTATIONS SCALE AND THE MIXED STANDARD SCALE: A COMPARATIVE EVALUATION

A Thesis

Presented to

the Faculty of the Department of Psychology

Western Kentucky University

Bowling Green, Kentucky

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Deborah J. Boniske

October 1979

AUTHORIZATION FOR USE OF THESIS

Permission is hereby

☒ granted to the Western Kentucky University Library to make, or allow to be made photocopies, microfilm or other copies of this thesis for appropriate research or scholarly purposes.

□ reserved to the author for the making of any copies of this thesis except for brief sections for research or scholarly purposes.

Signed Deborah J Boniske

Date 12-3-79

Please place an "X" in the appropriate box.

This form will be filed with the original of the thesis and will control future use of the thesis.

THE OPERATIONAL EFFECTIVENESS OF THE BEHAVIORAL EXPECTATIONS SCALE AND THE MIXED STANDARD SCALE: A COMPARATIVE EVALUATION

Recommended April 23, 1979
(Date)

Raymond M Mendel
Director of Thesis

Lawrence M. Hanser

Thomas Wm. Madron

Approved December 21, 1979
(Date)

Elmer Gray
Dean of the Graduate College

ACKNOWLEDGEMENTS

This study was made possible by the contribution of some very special people without whose help this study could not have been completed.

To Dr. Larry Hanser I would like to express appreciation for instruction and advice during a difficult phase of the computer analysis.

To Dr. Tom Madron and Carolyn Marks special thanks are due for the many hours of help and encouragement from the initial development of the data file to the completion of the computer analysis phase of this study.

Dr. Raymond Mendel, chairman of my thesis committee, deserves a special thanks. For his initiation of the study and his constructive criticism and suggestions during the review of rough drafts I wish to express my gratitude and appreciation.

Additionally, for suggesting the special analysis for interrater reliability, thanks to Dr. Leroy Wolins.

A special thanks goes to Danny Bean, Don Currie, and Tony Montebello for help during the data collection phase.

I want to give the final thanks to my parents. It was their continual support and encouragement that made this possible.

TABLE OF CONTENTS

LIST OF TABLES

THE BEHAVIORAL EXPECTATION SCALE AND THE MIXED STANDARD SCALE: A COMPARATIVE EVALUATION

D. Boniske October, 1979

Directed by: R. Mendel, L. Hanser, and T. Madron

Performance evaluations were obtained on firefighters in a large Midwestern City. The evaluations were conducted through utilization of two different scale types (The Behavioral Expectation Scale and The Mixed Standard Scale). These evaluations were obtained in order to test the hypotheses that the MSS was psychometrically superior to the BES in the reduction of halo and leniency error and that the MSS was also the better scale type in terms of producing higher interrater reliability. Leniency error (in both the absolute and comparative sense) was examined by conducting a series of T-tests. Halo error was investigated by a comparison of the means of the dimensions from each scale. The technique used to assess interrater reliability involved estimating the reliability of the differences in the shape and level of performance profiles of firefighters. The results showed that the first hypothesis, which proposed that use of the MSS produced less halo and leniency error than did use of the BES, was supported; however, the second hypothesis, which proposed that use of the MSS produces higher levels of interrater reliability, was not supported.

The Operational Effectiveness of The Behavioral Expectations Scale and The Mixed Standard Scale: A Comparative Evaluation

Effective and efficient utilization of manpower is a prime concern of all private and public organizations. For proper utilization of personnel, it is necessary that the organization identify its employee's strengths and weaknesses. Thus, a key component in the process of identifying and developing an organization's human resources is the periodic performance appraisal. Through these evaluations employees can be selected for promotions, and training needs can be identified. Additionally, through utilization of periodic evaluations an employee may be given feedback on his overall performance focusing on strengths and/or weaknesses. Also, since performance appraisals can be used to identify training needs, programs to strengthen the skills of an individual may be developed.

There are three fundamentally different types of commonly used performance evaluations: production measures, (e.g., number of units produced), personnel data measures, (e.g., absences, grievances, accidents), and judgemental measures (e.g., supervisory or peer ratings). In order to investigate which performance type was used most frequently, Blum and Naylor (1968) randomly selected 50 articles from the Journal of Applied Psychology between the years of 1960 and 1965. While perhaps an imprecise index of organizational usage, "Their results indicated a tendency for judgemental criteria to be used more of ten than either personnel or production data" (p. 197), and it is these judgemental measures which are the focus of this paper.

Problem

Over the years numerous attempts have been made to construct better types of performance rating scales, scales that are freer from the errors typically associated with ratings of performance. The specific problem focused on in this paper is that of determining which of two types of performance rating formats, the Behavioral Expectation Scale (BES) or the Mixed Standard Scale (MSS), is more psychometrically defensible. The determination is made through comparison of these two scale types on several important psychometric properties. These psychometric scale properties are halo error, leniency error, and interrater reliability. This comparison of scale types should shed light on the relative effectiveness of each scale type in minimizing these problems, problems the extent of which control the pragmatic value of judgemental performance ratings.

In the following discussion the first topic dealt with is the very nature of, and the developmental procedures utilized in, the construction of the two types of performance rating scales. The scale construction is followed by an explanation of the psychometric properties which we shall use as our measures of scale effectiveness. Next, the available literature comparing the relative effectiveness of rating scales is reviewed, focusing on the BES and the Summated Rating Scales (SRS). This evaluation is conducted because the SRS is the scale type most frequently compared to the BES in the literature. It is hoped that by focusing on the comparison of these two scale types light can be shed on the relative merit and effectiveness of the BES. The studies comparing the BES and the MSS are then examined, and finally an overall comparison of the two scales' effectiveness is made.

<u>The Behavioral Expectation Scale</u>

The Behavioral Expectation Scale's developmental procedure is more comprehensible once the logic behind the scale is explained. As Smith and Kendall (1963) have indicated two main objectives of the BES are to enhance the motivation of raters and to enable raters to do a better job of rating. These objectives are facilitated by the developmental practices utilized by the BES. For instance, the characteristics of the BES which enhance the raters' motivation include rater participation in scale development and the face validity of the final scales. The use of critical incidents of behavior, retention of rater terminology, and the retranslation process produce scales which should enable raters to do a better job in performance evaluation.

Certain characteristics inherent in the BES tend to deal with problems frequently encountered when dealing with other types of performance evaluations. As noted by Smith and Kendall (1963), the developers of the scale, "ratings from different raters in different situations should be really equivalent since they are almost always treated as if they were so . . . (I)nterpretation of the rating must not deviate too widely from rater to rater or occasion to occasion, either in level (evaluation) or in dimension (trait, situational characteristic, job demand, temporal requirements, etc.)" (p. 149). Additionally, they have pointed out that a problem in many performance evaluations has frequently occurred when psychologists who are developing the rating scales "have tended to impose their own values, interpretations and beliefs about behavior upon the raters" (Smith and Kendall, 1963, p. 149). The prescribed developmental procedures used in the construction of the BES deal with these problems and enhance the two main purposes of the BES (e.g., increased rater motivation, and enabling raters to do a good job).

Participation of supervisors and/or incumbents (raters) and retention of the participants' own terminology is an integral part of the scale development process. Through the retention of the raters' own terminology and their participation in all phases of the scale development more meaningful scale development is facilitated (utilization of the raters' own language reduces scale ambiguity). At the same time, the raters are "sold" on the rating scales through participation in their development and this in turn leads to increased motivation to use the scales by instilling in the participant a feeling of "ownership." In addition to selling the raters on the use of the scale, raters must be honest and careful in completing the ratings, and their honesty and carefulness are also facilitated by the fact that these scales have face validity (Smith & Kendall, 1963, p. 149).

A final advantage of the scale is its use of actual critical performance incidents as scale point anchors. Critical incidents are specific examples of effective or ineffective job performance behavior. To form the BES, critical incidents are identified in sessions with potential raters and are then reworded into an "expected behavior" format. Rewording the critical incidents into statements of expected behavior makes the rating task easier since many times critical incidents do not occur frequently enough on the job. The use of expected behavior statements gives the rater an opportunity to evaluate a given ratee's performance in terms of how the rater would "expect" that ratee to behave on a given dimension in a certain situation. That is, the use of expected behavior statements as anchors at the various performance levels of each dimension reduces the ambiguity present in many performance ratings, since these anchors give the rater a more job related perspective from which to make performance evaluations.

The BES developmental procedure is an iterative process which involves potential raters in all phases of development. The job under consideration is broken down into its major components or dimensions in group sessions composed of job supervisors and/or incumbents. The major dimensions are formally defined by the participants. Critical incidents of performance are then collected and each is assigned to its corresponding performance dimension. The critical incidents for each dimension are then subjected to a retranslation process. The retranslation is conducted by having a separate group of raters (those who did not participate in the original generation of dimensions or the assignment of critical incidents) take each critical incident and independently designate which dimension the critical incident exemplifies. Critical incidents are retained if there is a general consensus among raters as to which dimension they belong, and the dimensions are retained if the critical incidents are consistently reassigned to the dimension for which they were developed (Smith & Kendall, 1963, p. 152). A review of the literature indicates that different percentages of agreement among raters, usually ranging from 50% to 80%, have been utilized as the criterion for the retention of a given critical incident for a given dimension.

Incidents which are retained are then scaled, in terms of their relative effectiveness. The scaling is accomplished by asking a group of potential raters to independently designate the level of performance they feel a given critical incident represents by assigning the critical incident to a certain level on the dimension. Usually, a Likert rating format is used here. Those incidents that have been consistently reassigned to a dimension at a certain level (i.e., have a low standard deviation) are retained for the final scale. As stated by Borman and Dunnette (1975), "The net effect of using this scaling method should be decreased leniency error (because levels of performance are defined better), and higher interrater agreement (because raters are likely to be more cooperative and attentive to the rating task)" (p. 561). To form the final BES each dimension of the given job is graphically displayed on a separate page with a dimension definition stated at the top. That is, the final BES is a set of graphic scales with behavioral anchors that are stated in the rater's own terminology with each scale hopefully having a common and clear meaning for all potential raters (see Appendix A for example).

A final point for consideration is how the rating is obtained on the scale. The response format utilized by the BES is one that is logical and straightforward. Each rater must evaluate a given ratee's performance on each of the separate dimensions. The rating is completed by having the rater indicate the level on each scale which is most representative of the level at which a given ratee would be expected to perform on that given dimension. That is, the rater must respond once on each scale (dimension), with the critical incidents (rephrased to expected behavior)--which are used as anchors on the BES--giving the rater a frame of reference on which to base his rating of a given ratee on that one scale.

The Mixed Standard Scale

The Mixed Standard Scale (MSS), developed in 1965 by Blanz and Ghiselli, has been proposed as an alternative rating procedure designed to reduce several common rating errors. The developmental procedures of the MSS closely parallel those utilized in the formation of the BES. However, the format of the scale that results from this process is different, as is the response required of the rater.

To form the MSS, three critical incidents similar to those used on the BES to indicate high, medium, and low performance (on each scale) are collected for each performance dimension (see Appendix B for example). The three statements from each dimension which represent three levels of favorableness are then combined and arranged in random order (i.e., 7 dimensions x 3 statements = 21 items). Randomly arranging the statements "reduces the possibility that the rater will be able to form a clear picture of any order-of-merit set of descriptions for each characteristic being rated" (Blanz & Ghiselli, 1972, p. 187). These three statements from all the dimensions are randomly arranged in order to reduce the possibility that the raters become preoccupied with the order of merit of the incidents rather than focusing on the ratee's actual job performance in relation to each statement. Thus, halo error and leniency error are presumably lowered. Further, "the mixing provides a means for examining the dependability and reliability of the ratings, for it permits the ratings to be examined in light of consistency or logic of the three ratings on each dimension" (Blanz & Ghiselli, 1972, p. 187).

After the statements have been randomly ordered the rater evaluates each ratee as "worse than," "equal to," or "better than" each of the three statements (anchors). That is, the rater must respond once to each statement for a given ratee. Thus, a rater using the BES with seven dimensions each containing nine anchors would only rate each scale once (seven ratings in total), while a MSS developed to evaluate a ratee in the same position would result in a scale with twenty-one statements (three ratings on each of the seven dimensions) to be evaluated by a given rater.

Blanz and Ghiselli (1972) developed a 7-point scoring system to be employed in the evaluation of the rating of the three statements for each dimension. Their scoring system has a predetermined score for each possible combination of ratings that a given rater could assign a given ratee on the three dimensional statements. The scale is sensitive to illogical (as well as logical) rater responses. For example, if a given rater indicated that a ratee was worse than the lowest statement on a dimension and better than the middle and highest statement, the ratings would be indicative of an illogical response pattern and would receive a lower score than if the rater had indicated that the ratee was better than all three statements on the dimension.

The MSS has a number of unique advantages. For instance, the scoring system of the MSS allows the ratings on a dimension by a given rater to be checked for internal consistency. If a rater indicates that a given ratee is better than a superior behavior statement, it is logical that the rater should also indicate that the ratee is better than a statement (on the same dimension) indicative of poorer behavior. Furthermore, due to the MSS's format, examination can also be made of errors made in rating certain ratees and the number of errors made on a given dimension (trait). The above can be useful in detecting inconsistency in particular raters, ratees, and even particular dimensions. That is, by examining raters' scores it is possible to detect poor raters who may be in need of rater training. Similarily, if a ratee is consistently rated illogically by the raters, it may be that the raters have not had an adequate chance to observe the given ratee on the performance dimensions. Additionally, by looking at the ratings and the number of illogical responses on each dimension it may be possible to detect dimension statements that are ambiguous and need to be reworded. Finally, a further advantage of the MSS's scoring system, derived by having three ratings on each dimension, is an expected increased scale reliability. Instead of a single response on each dimension as obtained on the BES the MSS had three reponses on each dimension, and more reponses should lead to increased reliability (Nunnally, 1974, pp. 192-194).

Psychometric Properties

In dealing with judgemental rating systems it is necessary to realize and examine potential sources of rating error. Since by their nature judgemental ratings are subjective, they are susceptible to a number of common rating errors. Among the more prevalent rating problems are leniency error, halo error, and lack of interrater reliability. A problem arises when use of a given scale type produces such errors in that these errors reduce the validity of the ratings.

Leniency

Leniency error occurs when raters consistently give ratings substantially above (or below) the midpoint of the scale. This tendency to give all ratees more (or less) favorable ratings causes skewness and frequently produces a distorted picture of "true" performance. This type of error causes ratings to be "bunched together" and reduces the differentiation among poor and superior ratees (range restriction of ratings). Therefore, actual differences in performance between an average and an above average employee may be obscured when this rating error is prevalent. For example, a superior ratee should receive a high score (i.e., seven) on a seven point scale while a poor employee should receive a lower score (i.e., one). If leniency error is present and the poor employee receives an elevated score (i.e., four), then the meaningfulness of the evaluation is reduced. The problem is further compounded when ratings from different raters

Psychometric Properties

In dealing with judgemental rating systems it is necessary to realize and examine potential sources of rating error. Since by their nature judgemental ratings are subjective, they are susceptible to a number of common rating errors. Among the more prevalent rating problems are leniency error, halo error, and lack of interrater reliability. A problem arises when use of a given scale type produces such errors in that these errors reduce the validity of the ratings.

Leniency

Leniency error occurs when raters consistently give ratings substantially above (or below) the midpoint of the scale. This tendency to give all ratees more (or less) favorable ratings causes skewness and frequently produces a distorted picture of "true" performance. This type of error causes ratings to be "bunched together" and reduces the differentiation among poor and superior ratees (range restriction of ratings). Therefore, actual differences in performance between an average and an above average employee may be obscured when this rating error is prevalent. For example, a superior ratee should receive a high score (i.e., seven) on a seven point scale while a poor employee should receive a lower score (i.e., one). If leniency error is present and the poor employee receives an elevated score (i.e., four), then the meaningfulness of the evaluation is reduced. The problem is further compounded when ratings from different raters with differing levels of leniency are directly compared by a third party. For instance, problems could occur when a given group of raters (although adequate performers) were not leniently rated by their supervisor. Then, a comparison made by a superior between that group and another group whose supervisor had rated leniently would tend to make the first group's performance seem inadequate. This lower evaluation resulting from a comparison of the groups would be due to the higher levels of leniency associated with the initial supervisor's evaluations of his subordinate's performance and not the first group's actual performance.

Halo

Halo error has been defined as the tendency for a rater to evaluate a given ratee at a similar level on all dimensions based on the evaluation of or generalization from only one dimension or a global impression. Such evaluation results in high intercorrelations among ratings for supposedly different characteristics or behaviors. Halo represents the failure of the rater to differentiate among different characteristics of performance. Lack of differentiation among performance dimensions in evaluation of a given ratee is a serious problem. The fact that given ratee's performance is evaluated as superior on one dimension provides little basis to infer that his performance is at the same level on all the other dimensions. Thus, a ratee who performs well on a given dimension may perform poorly on other dimensions and should be evaluated as such. As previously stated, a major purpose of performance evaluations is the identification of an employee's relative strengths and weaknesses. When halo error occurs it creates a problem in that the performance evaluation is muddled. That is, through performance evaluation areas can be identified where employees are lacking and their development can be facilitated. However, when halo error occurs it obscures these weaknesses and reduces the chances for the development of employees. For this reason, it is imperative that halo error be minimized.

Interrater Reliability

The type of reliability dealt with in this paper is interrater reliability. This type of reliability is determined by assessing the degree to which raters agree on their ratings of different ratees. This study investigates the effectiveness of two scale types once they have actually been used in an evaluative situation, and it is for this reason that interrater reliability is assessed rather than scale reliability.

BES Compared to SRS

The MSS is a relatively new scale type; therefore, few studies have been conducted comparing the MSS to the BES. The vast majority of psychometric information pertaining to the Behavioral Expectation Scales (BES) is derived from studies which compare it to Summated Rating Scales (SRS). Therefore, since the SRS is the most widely used alternative rating scale, it should be useful to review the literature comparing the BES with SRS. By reviewing this literature it is possible to see the relative effectiveness of the BES in dealing with certain psychometric properties. The psychometric considerations on which the two scale types are evaluated include leniency error, halo error, and interrater reliability.

Several researchers found less leniency when the BES was used as compared to the SRS (Borman & Dunnette, 1975; Burnaska & Hollman, 1974; and Campbell, Dunnette, Arvey, & Hellervik, 1973). Another group of researchers found no significant differences in leniency when using the BES and the SRS (Bernardin, 1977; Friedman & Cornelius, 1976; and Keaveny & McGann, 1975). Still, others found that use of the BES produces more leniency then the SRS (Bernardin, Alvares, & Cranny, 1976; and Borman & Vallon, 1974).

When examining halo error associated with the two scale formats several researchers found the BES reduced halo error in comparison to the SRS (Borman & Dunnette, 1975; Campbell, et. al., Friedman & Cornelius, 1976; and Keaveny & McGann, 1975). In three other studies halo error was found to be virtually equivalent for the two scale types (Bernardin, Alvares, & Cranny, 1976; Bernardin, 1977; Borman & Vallon, 1974). A final study found halo error present in all ratings on the two scales (Burnaska & Hollman, 1974).

Before reviewing the literature on reliability it is necessary to point out that reliabilities obtained on the BES, which are computed in the developmental phase (scale reliabilities), are not an adequate means by which to evaluate the "operational effectiveness" of the BES. These scale reliabilities are computed by dividing the group of raters who scaled the anchors into two equal groups, computing a mean value for each anchor on each dimension, and then correlating the two sets of means from each group (Borman & Vallon, 1974). Reliability determined in this procedure merely reflects how reliably raters rank statements. This type of reliability provides no information about the scale reliabilities when actually used for rating purposes. Therefore, of the possible types of reliability, reliabilities obtained in the scale developmental phase are the least desirable for determining the operational effectiveness of the BES. Since reliabilities obtained in the developmental phase are deemed inappropriate, it is advisable to investigate the reliabilities obtained when the BES has actually been applied in an evaluative situation. That which should be investigated is the operational effectiveness of a scale type rather than the scale development effectiveness. Operational effectiveness is determined through actual scale application in an evaluative situation (the scales are actually used to evaluate ratees). An appropriate index of operational effectiveness is interrater reliability. As pointed out by Borman and Vallon (1974), and Zedeck and Baker (1972), interrater reliability is superior to scale reliabilities in judging the "usefulness" of a rating format. Interrater reliability can be defined as the agreement between two or more raters using a given scale type for a given ratee(s). Only literature which deals with studies of the operational interrater reliability of scales will here be reviewed.

Several researchers found that interrater reliability was greater when using the BES as compared to the SRS, although the difference was slight (Borman & Dunnette, 1975; Borman & Vallon, 1974). No significant differences in interrater reliability were obtained by Bernardin (1977) when using the BES compared to the SRS, and in a recent study the SRS was found to produce higher interrater reliability (Bernardin, et. al., 1976).

There are a number of possible explanations for the different results in the studies dealing with leniency, halo, and interrater reliability. Among the more prevalent reasons offered for these differences are differences in scale development procedures, amount of supervisor and/or incumbent participation in development, the level at which critical incidents are retained (or eliminated) and scaled for effectiveness, scale familiarity, whether or not raters received training on the different types of rating errors, and differences in the sample of raters or the type of person being rated. These differences are explored more fully in the following sections.

Developmental differences

The first three explanations deal explicitly with the development of the scales and the resulting format. As noted by Bernardin, Alvares, and Cranny (1976), "The methods of constructing scales are as important as the finished product . . . Thus, in comparing rating formats, the method of scale development should be examined first. If comparable effort has not been exercised to insure equally rigorous scales, the implications of the results are hopelessly confounded" (p. 569). In the above studies the SRS was developed in a number of different ways. One procedure used the definitions developed for dimensions on the BES (Campbell, Dunnette, Arvey & Hellervik, 1973). The researchers broke the BES's scale dimensions down into their major components and then scaled each of these with a Likert type format. Another procedure employing the same dimensions and definitions as the BES anchored the scaled with verbal descriptions (Keaveny & McGann, 1975). Yet, another procedure used performance dimensions that were different from those used on the BES (Burnaska & Hollman, 1975). A final procedure used anchors that were retained after the retranslation process had occurred on the BES (Bernardin, 1978).

It is reasonable to expect that these different methods of scale development could cause differences in leniency, halo, and interrater reliability. Some of the SRS developmental techniques produce scales which approximate the BES (i.e., scales constructed from the same dimensions and dimension definitions as those used on the BES), thereby making it easier to rate a given ratee. That is, scales developed from the dimensions and dimension definitions on the BES give the rater a framework helpful in his evaluation of a given ratee, and may result in less leniency error and halo error. However, scales developed from just the behavioral anchors (that survived the retranslation process and were scaled for effectiveness) and which use a seven point response format (Bernardin, 1977, p. 423) do not indicate to the rater which dimension they represent nor do they give an indication of the anchor's level of favorableness. Furthermore, the above could produce some scales with more anchors than others (more anchors per scale dimension). Given a set of dimensions it is desirable to have at least five (or more) anchors from each dimension survive the retranslation process. However, if they did not, certain dimensions could presumably have a larger number of anchors and these anchors could be representative of different levels of scale favorableness. On one scale anchors representative of favorable performance levels may have survived the retranslation process while anchors representative of lower levels of performance may have been the ones to survive the retranslation process from another dimension. Thus, differences in leniency error and halo error could be due to the ambiguity created by the different number of anchors for each dimension (on each scale) and the different levels of favorableness represented by the anchors. Furthermore, the lack of a frame of reference at certain points on the dimensions may result in differences in interrater reliability.

Bernardin, LaShells, Smith, and Alvares (1976) pointed out that differences in results obtained using the BES could be traced to the developmental procedure used as well as format differences. They found that one difference in the developmental procedure was the level of agreement between raters, which was utilized as a criterion for the retention of critical incidents. The format differences included the use of continuous versus noncontinuous scales and the "use or nonuse of dimension clarification statements at anchor points on the scale" (p. 75).

From reviewing the above literature it was found that not only did retention level for critical incidents differ (50% to 80% agreement in assignment to a given dimension for retention), but that the standard deviation of judgements of effectiveness of critical incidents used as a criterion for retaining anchors also varied. The larger the standard deviations are (in regard to assigning a critical incident to a scale at a certain level) the more ambiguity is present in the final scale. When agreement between potential raters is low (large standard deviation) as to the effectiveness of a given anchor the resulting scale may have anchors that overlap, thus making differentiation among them difficult and increasing leniency error and reducing interrater agreement. Also, if the percentage agreement for retranslation is low, anchors may be too similar across dimensions, and could increase the difficulty associated with differentiating among dimensions and agreeing on ratings. The reduction in differentiation would increase halo error and reduce interrater reliability thereby reducing the meaningfulness of the resulting ratings.

A final difference in developmental procedures occurs when the same set of judges develops and retranslates the dimensions (Campbell et. al., 1973; and Keaveny & McGann, 1975). Use of the same judges may cause an increase in halo error due to the fact that incidents have not properly been reassigned, or that the dimensions are too closely related.

Format Differences

It was also noted from reviewing the literature that studies differed in the scale format they employed. The differences included continuous versus noncontinuous scales and differences in response format (seven versus nine point scales). For example, on a continuous BES scale (with seven anchors placed at various points along a continuum) the rater may respond at any point along the scale from 1 to 7, while on a noncontinuous scale anchors are placed at equal intervals (from 1 to 7) on the scale and the rater must respond to one of the seven points. Leniency error may be increased on the noncontinuous scales due to the restriction in response options available to the rater. That is, on the noncontinuous scale the rater is forced to choose between a limited number of responses thereby reducing the ability to differentiate among employees.

Participation

Another difference in these studies is whether or not potential raters participated in the development of the rating scales. As previously reported by Friedman and Cornelius (1976) participation in scale development can lead to increased scale precision regardless of the scale format (BES or SRS). "That is, regardless of which scale format was used, subjects who had participated in scale development provided ratings that were psychometrically superior" (p. 215). They also reported that participation of potential raters later led to increased rater motivation and conscientiousness on the rating task. Therefore, those scales which benefited from participation in the developmental phases may have reduced leniency error and halo error due to the raters increased motivation to rate carefully. Participation may also lead to higher interrater reliability due to an increase in the understanding of the scale type and increases in agreement on the definition of performance dimensions.

Scale Familiarity

The familiarity the raters have with the two scale types may also influence leniency error, halo error, and the degree of interrater reliability obtained. Frequently, simpler types of rating scales (SRS) have been used. Often when a rater is more familiar with a given scale type he has less difficulty in completing the rating. That is, when a rating scale is employed that the rater is unfamiliar with the rater must take time to understand the scale before he can rate a given ratee. Therefore, simply familiarity with a scale type may affect the obtained leniency error, halo erro, and interrater reliability. Thus, these psychometric properties would be modified due to an external consideration rather than an actual rating scale property.

Rater Training

Differences in the studies (on all three of the psychometric properties) may also have been related to the amount of training the raters received on the different rating problems. Raters in a given study may have had training on the types of rating errors typically associated with judgmental ratings, while raters in a second study received no training on the errors. If raters in one study received training on these errors and raters in another did not, it would be logical to expect the resulting ratings to differ (Borman, 1975; Bernardin, 1978; and Latham, Wexley, & Pursell, 1975).

Differences in Raters and Ratees

A final reason for the obtained differences may be due to differences in the raters and ratees chosen for the studies. The studies employed raters from a number of different occupations (e.g., head nurses, teachers, first line - and second line supervisors). Additionally, the ratees also came from equally diversified areas (students, nurses, managers). Due to the different backgrounds of the raters and ratees, it is plausible to expect differences in the psychometric properties with which this study deals.

From the literature comparing the BES to the SRS, a number of conclusions can be drawn. First of all, insofar as leniency error is concerned there is little evidence that either scale format is superior. Developmental differences, as well as others (e.g., format differences), may be reasons for this inconclusiveness. Three studies found leniency was reduced when using the BES, while three other studies reported leniency to be present and relatively the same on both of the scale types (BES and SRS), and the last two studies reviewed found leniency error to be higher when the BES was used, as compared to the SRS.

The literature which compares the BES and the SRS does lend some support to the notion that the BES is superior in reducing halo error.

However, the evidence is not by any means conclusive. Four studies showed that halo error was reduced when the BES was used. Three other studies reviewed found halo error to be relatively equivalent for the two scale types. The final study reviewed indicated that the BES produced more halo error than did the SRS.

For interrater reliability the results are also inconsistent. Two of the four studies reviewed found interrater reliability was higher when using the BES. The other two studies found interrater reliability to be the same as that obtained when using the SRS and lower than that obtained by use of the SRS, respectively.

Although the results appear inconsistent from the review of the literature, there is slightly more evidence supporting the effectiveness of the BES than the SRS. Even though the psychometric evidence pertaining to the use of the two scale types only marginally favors the BES, there are a number of non-psychometric advantages inherent in the BES (stemming from the format itself and the developmental process) which recommend its use. These advantages include the identification of areas for training, increased motivation of raters, participant learning, reducing role ambiguity, etc. Additionally, advantages have also been cited by Blood (1974). He has pointed out that the BES may be used to "extend the domain of evaluated performance" (p. 514).

In the final analysis then, since the psychometric considerations marginally favor the BES over the SRS--and there seem to be substantial (although not well documented) ancillary benefits stemming from the development of the BES--the BES appears to be superior to the SRS as a performance rating strategy.

The Mixed Standard Scale

The MSS seems to have a number of advantages over conventional rating scales. Through the scale's format and scoring system the MSS is able to deal with conventional rating errors (halo error, leniency error, and lack of interrater reliability). For instance, when leniency error is detected, statements may be rescaled to a higher level to increase differentiation among ratees (statements may have been examples of behavior that were too low and by rescaling the examples to a higher level or generating new examples the rater has a clearer frame of reference within which to make an evaluation). Also, the scale format of the MSS (randomly arranged anchors) presumably reduces the chances that a given rater's rating will be affected by halo error since the scarmbling of the anchors presumably disguises somewhat the dimension to which a statement pertains. As previously noted, the MSS also allows for more than one rating on each trait, and thus the MSS should produce relatively higher reliabilities. Another advantage of the MSS evolves from its scoring procedure. As mentioned above, the scoring system allows for ambiguous statements to be recognized and rewritten by noting a rater's logical error scores and for rater inconsistency and scale dimension ambiguity to be examined by checking a rater's score from ratee to ratee or dimension to dimension.

Due to the recent development of the Mixed Standard Scale (i.e. 1965), relatively few studies have been conducted using the MSS, and only two of these compare the MSS and the BES on the psychometric properties with which this study deals. Therefore, relevant studies which deal with other psychometric properties will also be reviewed in order that the effectiveness of the MSS may be examined.

The original MSS was developed in Finland by Blanz (1965) and was used in the evaluation of low level employees on several jobs. The studies, reported below, used developmental procedures similar to those used by Blanz and were conducted in the United States.

The first of these studies was conducted by Blanz and Ghiselli (1972) to determine whether the MSS could be applied at a different organizational level (managers) and whether the scale type was applicable in a cultural setting different from that in which it was developed. The results showed that halo error and leniency error were low (in an absolute sense) when the MSS was used. Also, due to the scaling method and the scoring system used they were able to detect the consistency with which a given rater rated. Additionally, they deemed the evaluation of rater consistency an appropriate measure of reliability.

In a study conducted by Arvey and Hoyle (1974) the MSS and the BES were both employed in order to determine the scale's relative effectiveness and to see if "good" versus "poor" raters could be differentiated by use of the two scale types. The results showed only a slight tendency for raters who rated poorly on one dimension to do so on another (p. 67). However, when examination was made of the multitrait-multimethod matrix both scales did exhibit relatively good convergent and discriminant validity. "Convergent validity is demonstrated when correlations between the same dimension using different rating methods are significantly different from zero and reasonably high . . . Discriminant validity may be defined as the extent to which a dimension or trait is differentiable from other dimensions" (p. 64).

Finley, Osburn, Dubin, and Jeannerett (1977) investigated the effectiveness of scales with specific anchors and scales with general anchors, and whether or not differences in the scale format affected the ratings. For their purposes they used two different scale formats--one of which resembled a BES format (obvious presentation of anchors), while a second more closely approximated a MSS (presentation of anchors was mixed). There was little evidence to support the notions that either scale type or specificity of anchors (obvious-BES versus disguised-MSS) was better in reducing leniency error or halo error. However, the behaviorally anchored scale was superior to the scale which resembled the MSS in producing higher interrater reliability, regardless of whether general or specific anchors were employed.

Though the Mixed Standard Scale has several theoretical advantages over the Behavioral Expectation Scale, it is apparent from the review of the literature that the results of comparative studies though few in number are mixed. That is, although the underlying premises of the MSS have merit (e.g., disguised scale obviousness), the results of comparative studies are perplexing. An example is the lower levels of interrater reliability exhibited on the MSS when compared with the BES. However, this phenomena may be due in part to other properties of the scale (i.e., reduction of halo and leniency error). Therefore, the purpose of this study is to further examine these discrepancies. Based on the foregoing literature, two hypotheses are offered. The first hypothesis is that the MSS will have less leniency and halo error than the BES. Moreover, due to the larger number of items on the MSS and despite some prior conflicting evidence higher interrater reliability is hypothesized for the MSS.

METHOD

Overview

The general strategy employed in this investigation involved the development of two rating scale formats (BES and MSS). Once the scale formats were developed, fire captains (raters) were asked to rate all the fire fighters (ratees) with whose performance they felt they were sufficiently acquainted. The two scales were administered at approximately the same time; however, the MSS was administered first due to its format (mixed presentation of anchors from all the dimensions). It was believed that if the BES was administered before the MSS the anchors order-of-merit would be apparent to the raters. All of the fire fighters were rated using both types of rating scales. Afterward the rating scales were compared for leniency, halo, and inter-rater reliability.

Raters-Ratees

The ratees and raters selected for this study were male fire fighters and fire captains in a large mid-western city. The ratees (N=74) were fire fighters who had been appointed following successful performance on an employment test, had succeeded in training, and had been on the job for at least six months. The raters were fire captains (N=58) who had at least six months contact with the fire fighters.

Rating Scales Development

Behavioral Expectation Scales

The BES was developed in a manner similar to that proposed by Smith and Kendall (1963). From a previous job analysis questionnaire a number of examples of effective and ineffective fire fighter performance had been elicited. An analysis of the examples resulted in seven major performance areas: learn and apply, judgement under stress, physical fitness, compatability, public relations, teamwork, and pride and dedication to career. These dimensions were presented to fire captains in a number of sessions, and there was general consensus that the identified dimensions were appropriate. In these sessions participants were asked to write specific behavioral examples of a positive and negative incident for each dimension. The incidents from a job analysis questionnaire and those generated by the sessions were combined and randomly arranged on another questionnaire along with the description of the seven performance dimensions. Independently, fire captains reassigned the incidents to dimensions. If an anchor (critical incident) was reassigned to the same dimension by 56% of the fire captains it was retained. Researchers and representatives of the fire captains and chiefs reviewed the retained incidents combining any duplications and assigning incidents to their respective dimensions. The regrouped incidents (at least twenty per dimension) sorted into their respective dimensions were given to groups of fire fighters and captains to assign a dimension effectiveness value of from 1 to 7 (low to high) using a Likert format. Incidents were retained if they satisfied the 65% criterion discussed above and if there was general agreement among captains as to the value of each incident (i.e., if the standard deviation of the rated effectiveness of each incident was less than 1.70). To counteract a problem which frequently occurs when using behaviorally based ratings--insufficient incidents anchoring the midpoint or average performance level for each dimension--the researchers generated several incidents illustrative of average performance and again subjected the incidents to retranslation. Incidents consistently retranslated and possessing low standard deviations were retained. All incidents were then edited to the form "This fire fighter could be expected to . . . " and were placed at their assigned scale level from low to high effectiveness on the graphic scale. Finally, a global performance rating item was added to the end of the BES form.

Mixed Standard Scale

The Mixed Standard Scale (MSS) was developed directly from the BES. For each performance dimension three incidents were selected to form the MSS. The anchors chosen were those previously shown to have a small effectiveness rating standard deviation, in the developmental phase of the BES. Based on the results of the previously administered BES it was evident that raters were too lenient in their evaluations. Therefore, in this study, items were selected for MSS scale inclusion that had higher rated effectiveness values. By so doing it was anticipated that a more normal distribution of ratings would result (i.e., less leniency). The three anchors (incidents) chosen for each dimension from the BES were edited from expectations to actual incidents of behavior to form the MSS. The resulting MSS had (7 traits by 3 statements per trait) twenty-one statements.

The scoring procedure developed by Blanz and Ghiselli was used

(This scoring procedure is presented in Appendix C). Once the evaluation of the ratees' performance had been completed, their scoring system was used to convert the three statements on each dimension to a numeric score. For purposes of expediency, a computer program by McPhail and Dickenson (1977) was used to convert the ratings to final scale scores. The scoring included predetermined numeric scores for logical as well as illogical responses. An additional advantage of the scoring procedure was the fact that it allowed for checks on rater consistency, the "number of errors per scale, and the number of errors per ratee" (Blanz & Ghiselli, 1973, p. 88).

Procedure

The fire captains were instructed to rate all the fire fighters with whom they had been in contact for at least six months. If the fire captains felt they had not had sufficient opportunity to observe the performance of a given fire fighter they were given the option of not rating that individual. As part of a test validation study the BES had been previously developed. Fire captains were asked to rate any fire fighters with whom they had contact and about whom they felt they had adequately observed performance on all dimensions. All fire fighters who were rated by at least two fire captains were selected for the study. Where more than two raters had rated a particular fire fighter, two raters were randomly selected and arbitrarily designated as rater one or rater two for purposes of analysis. The BES and the MSS were administered the same day. However, the MSS was administered first.

Analysis: Leniency Error

Leniency error was defined as a significant shift in mean ratings from the mid-point of the scale in either the positive or negative direction. The average performance of ratees is expected to be near the midpoint of the scale and should be reflected in the actual ratings. Two different methods were used to assess the amount of leniency error present in the two types of ratings scales.

The first test for leniency used the hypothesized "ideal" scale mean of 4.0 as a basis for comparison. (The BES and MSS were based on a seven point scale and on a seven point scale the hypothesized mean is 4.0). This analysis was performed to determine the degree of "absolute" leniency error (whether the means of the dimensions on each scale differed significantly from the hypothesized scale mean of 4.0). Each scale dimension mean was tested against the hypothesized scale mean through a series of t-tests to determine whether the obtained mean was significantly different from the theoretical midpoint of 4.0.

The second analysis was conducted to determine whether there was a significant difference between the mean ratings on the corresponding dimensions for the two scales. That is, the amount of relative leniency error was assessed by conducting a series of t-tests. In this instance, relative leniency is defined as the extent to which the mean score on the MSS is different from the mean score on the BES. The mean rating on each dimension of the BES was compared with the same dimension on the MSS. Fourteen separate t-tests (seven for BES and MSS rater one, and seven for BES and MSS rater two) were computed.

Halo Error

As previously noted, halo error is a tendency for raters to evaluate a given ratee at the same level on all dimensions. In this study halo error is evaluated by intercorrelating dimensions within each scale type, converting the correlations to Fisher's Z, calculating the average Z between each of the dimensions and all other dimensions for each scale type, and then converting the Z back to r to obtain the average intercorrelation for each dimension with all other dimensions for each dimension and scale type (Snedecor & Cochran, 1967, p. 185-187). The logic behind this analysis is that the higher the resulting correlation coefficients the less well the rater has discriminated among different dimensions and, therefore, the higher the halo error. For this study, a test of significance was deemed inappropriate due to a problem in the way the ratings were obtained and was therefore not conducted. Since some raters were used across more than one ratee the assumptions underlying a test of significance (independence of ratings) would have been violated.

Interrater Reliability

As mentioned above, to assess the reliability of the two scale types the operational reliability, as opposed to developmental reliability, should be investigated. Due to the manner in which the ratees were rated the traditional intercorrelation techniques were deemed inappropriate. In this study some raters rated more than one ratee on both scale types and some raters only rated one ratee on both scales. In order to obtain two raters per ratee, of those who had evaluated a ratee on both scale types, two of the above ratees were randomly selected. These two raters were arbitrarily designated rater one and rater two; however, raters were not the same individuals across all ratees. An appropriate technique to use for assessing reliability under these design constraints involved estimating the reliability of the differences in the shape and level of performance profiles of firefighters.[1] This technique involved computation of multitrait-multimethod matrices (Campbell and Fiske, 1959) for each scale type (BES and MSS) where raters are considered methods and dimensions are treated as traits. Two types of reliability can be estimated using this approach. One type estimates the reliability of differences in the level of the profile (a profile is conceived of as the scores of any individual on the seven dimensions). A second type of estimate of reliability is the reliability of differences in the shapes of profiles between people when using a particular scale type. Computational formulas for each are shown:

1. Level

$$\frac{N\ (\text{average hetrotrait-hetromethod})}{1+(N-1)(\text{average hetrotrait-hetromethod})} \qquad (1)$$

N=number of raters (e.g., 2) x number of traits (e.g., 7)

2. Shape

$$\frac{N\ (\text{average difference (hetromethod-monotrait) - hetro hetro})}{N-1\ (\text{average difference (hetromethod-monotrait) - hetro hetro})} \qquad (2)$$

N=number of raters per ratee (e.g., 2)

The first formula above (Equation 1) produces a correlation which can be interpreted as reflecting the reliability of differences in the levels of profiles of different ratees (i.e., high versus low performance across dimensions). These correlations when inserted into a Spearman Brown formula estimate the extent to which the differences in level are reliable when all seven dimensions

[1] Appreciation is expressed to Dr. Leroy Wolins for suggesting this analysis.

are considered simultaneously. That is, the reliability of the differences in level of profile across the seven dimensions is assessed.

The second estimation of reliability is of a different nature (see Equation 2). In this instance, correlations are again utilized from the multitrait-multimethod matrix. Computations similar to those for the determination of the differences in reliability of level are calculated in order to estimate the reliability of the differences in the shapes of the profiles. The reliability of the differences in the shapes of profiles is estimated by looking at the average difference between the hetrotrait, hetromethod correlations and the correlations from the validity diagonal. Here, the validity diagonal is composed of correlations between ratings using different methods on the same trait across all subjects. The reliability of the shape of a profile is determined by utilizing the second formula above, and it involves the averaging of the validity diagonals for each scale type.

The above estimates of the reliability of the differences in level and shape are computed based on two raters. An estimate of the reliability of both level and shape is also computed for one rater. The same formulas (Equations 1 and 2) are utilized (with minor modifications). To estimate the reliability of the differences in the levels and shapes of profiles with just one rater the same correlations are obtained from the multitrait-multimethod matrix; however, substitutions are made in the Spearman Brown formula to reflect the fact that only one rater does the evaluation.

Results

Leniency

The first set of t-tests, which were conducted to measure the difference between the scale dimension mean and the hypothesized scale mean of 4.0 (absolute leniency), revealed that the MSS was superior to the BES in producing means that were closer to the hypothesized scale mean of 4.0. As reported in Table 1 the majority of the scale dimension means on the MSS were not significantly different from the "ideal" mean. Of the seven t-tests conducted for each scale type only two of the scale dimension means were significantly different on the MSS ($p.\leq.05$), while all of the scale dimension means on the BES departed significantly from the "ideal" scale mean of 4.0. Thus, there is clear evidence that there was substantial leniency error present when the BES was utilized, while little leniency error was present when the MSS was used.

The analysis conducted to determine relative leniency error present in the two scales revealed that there were indeed significant differences in leniency error between the two scales. As can be seen in Table 2 the scale dimension means of the MSS were significantly different from those reported for the BES. In all seven cases the scale dimension means of the ratings on the MSS were lower than those of the BES and in all seven cases the difference was significant beyond the .01 level.

Table 1

Mean Dimensions Compared to the "Ideal" Scale Mean

Dimension	Scale	Mean	T-Value
Learn and Apply	BES	4.729	3.806**
	MSS	3.865	1.310
Judgement Under Stress	BES	4.987	4.945**
	MSS	4.000	0
Compatability	BES	4.824	3.628**
	MSS	3.865	.883
Physical Fitness	BES	5.095	6.014**
	MSS	3.595	2.388*
Public Relations	BES	4.878	5.663**
	MSS	3.689	2.859**
Pride and Dedication to Career	BES	4.581	3.020**
	MSS	3.487	3.694**
Teamwork	BES	4.824	4.482**
	MSS	4.054	.341*

Number of Cases=74

*p.<.05
**p.<.01

Table 2

Mean Dimension Ratings - Leniency Error

Dimension	Scale	Mean	Standard Deviation	T-Value
Learn and Apply	BES	4.729	1.649	4.84*
	MSS	3.865		
Judgement Under Stress	BES	4.987	1.716	5.02*
	MSS	4.000	1.194	
Compatability	BES	4.824	1.954	4.51*
	MSS	3.865	1.317	
Physical Fitness	BES	5.095	1.563	7.83*
	MSS	3.595	1.461	
Public Relations	BES	4.878	1.334	7.93*
	MSS	3.689	.935	
Pride and Dedication to Career	BES	4.581	1.655	6.28*
	MSS	3.487	1.196	
Teamwork	BES	4.824	1.582	4.03*
	MSS	4.054	1.364	

Number of Cases=74

$*p.<.01$

Halo

The correlations computed to assess halo error revealed that the use of the MSS produced dimensions with lower mean intercorrelations. As is evident from reviewing Table 3, in all cases the mean intercorrelations of the scale dimensions were higher for the BES than they were on the MSS. From the table it is also apparent that the intercorrelations of the two scales on each dimension are different from one another. That is, it appears that the differentiation among dimensions was greater when the MSS was used. Thus, the use of the MSS appears to have decreased halo error.

Interrater Reliability

The results from the reliability analysis are mixed. From Table 4 it is apparent that the BES was associated with higher estimates of the reliability of the differences in level. Additionally, the reliability of differences in levels was higher when the BES was used as compared to the MSS for both one rater and two raters.

On the other hand, when the reliability of differences in shapes was investigated the results indicated (as shown in Tables 4 & 5) that use of both scale types (BES and MSS) produced low interrater reliability. However, the reliability of differences in shapes was slightly higher when the MSS was used (for both one and two raters).

Table 3

Halo Error
Mean Correlations Among Dimensions

Dimension	Scale	Mean Correlation
Learn and Apply	BES	.45
	MSS	.29
Judgement Under Stress	BES	.56
	MSS	.26
Compatability	BES	.54
	MSS	.30
Physical Fitness	BES	.48
	MSS	.21
Public Relations	BES	.40
	MSS	.26
Pride and Dedication to Career	BES	.56
	MSS	.34
Teamwork	BES	.63
	MSS	.40

Number of Cases=74

Table 4

Reliability Estimate - Two Raters

Type of Reliability	Scale	Reliability Coefficient
Level	BES	.79
	MSS	.17
Shape	BES	.17
	MSS	.25

Number of Cases=45

Table 5

Reliability Estimate - One Rater

Type of Reliability	Scale	Reliability Coefficient
Level	BES	.66
	MSS	.09
Shape	BES	.13
	MSS	.23

Number of Cases=45

Discussion

This study was conducted in order to determine which of two scale types (BES and MSS) was more free of leniency and halo error and which possessed greater interrater reliability. The analysis of the data produced moderate support for one of the hypotheses. From the results (Tables 1 through 3) it is apparent that the first hypothesis which was proposed (e.g., MSS is superior to BES in reducing leniency error and halo error) was confirmed. However, the second hypothesis (use of the MSS produces higher interrater reliability than does use of the BES) was obviously not supported.

A number of plausible explanations for the obtained results will be offered and discussed. This explanation will be followed by an explanation of the possible bearing the results have on areas for future research.

Leniency Error

In the present study the two different methods that were used to assess the amount of leniency error in the two types of rating scales revealed that the MSS was superior to the BES in the reduction of leniency error (both in terms of absolute and comparative leniency). The first method employed to detect leniency error compared the two scale types against the hypothesized "ideal" scale mean. From the results (Table 1) it is obvious that in the absolute sense the utilization of the BES produced significantly more leniency error than did the MSS. The second method, which also used a series of t-tests, compared the mean ratings on each scale dimension and found that use of the BES resulted in significantly higher levels of leniency than did the MSS.

These findings are not surprising in view of the developmental and format differences between the two scales. One of the most important explanations for the findings on leniency error stems from a developmental difference. Based on a prior administration of the BES with fire captains, leniency error was detected. For this reason, instead of choosing statements from the BES indicative of poor, average, and superior performance, the MSS was developed by selecting statements which were representative of higher levels of performance (i.e., moderately poor performance, moderately high performance, and superior performance). By selecting these statements to form the MSS it was posited that the raters would be forced to differentiate more clearly among ratee's performance. However, because the BES was not rescaled, as was the MSS, this lack of rescaling may be a possible explanation for the differing results obtained for leniency error when utilizing the two scale types.

A second explanation for the higher levels of leniency stems from the format differences between the scale types. The format of the BES graphically presents each dimension along with its respective statements, thereby revealing the statement's favorableness. However, on the MSS the statements are randomly arranged, thereby reducing the possibility that the raters may determine the statement's order-of-merit. That is, by removing the graphic cue as to the statement's favorableness, raters may be more apt to be conscientious in the rating task and the resulting ratings are apt to have decreased leniency error.

A final interpretation of the findings is derived from consideration of the familiarity the raters had with the two scale types. As mentioned above, the BES had been previously administered, so the majority of raters were somewhat familiar with the scale type. Conversely, the MSS was a scale type which was new to the raters so they may have attended more carefully and conscientiously to the content, and as a result, may have rated less leniently. Only further investigations addressing these alternative possibilities will permit more definitive statements of the causal relationships.

Halo Error

The findings reported for halo error revealed that the utilization of the MSS resulted in decreased halo error compared to the BES. The potential explanations for the reported results include the scales sequence of administration, the familiarity that the raters had with the two scale formats, differences in the reliability of the two scales, and the inherent superiority of the MSS.

The MSS, as explained above, was the first scale type to be administered. Due to the MSS's format, it was not deemed advisable to counterbalance the order of scale administration. It is possible, therefore, that the results reported for halo are an artifact of the sequence of the administration of the scales and not actually a property of the scales. In other words, because the MSS was administered first, the raters may have been more interested in and attentive to the rating task. However, by the time the raters had completed the MSS and began the evaluation using the BES the novelty of the rating task using a new scale may have vanished. Thus, the task may have been completed more rapidly and less conscientiously due to boredom and disinterest.

A second plausible explanation deals with the familiarity the raters had with the scale format and follows directly from an inherent characteristic of the MSS. As Blanz and Ghiselli (1972) have noted, halo error is decreased when "the ratings are not obviously made on a scale" (p. 186). Since the MSS was a scale type with which raters were unfamiliar, and because the statements order-of-merit was disguised by not having placed them on a graphic scale, it is possible that the raters were more conscientious and exhibited greater attentiveness in the rating task.

The amount of halo error present in the two scale types is directly related to the reported reliability. The utilization of the MSS produced scales with virtually no interrater reliability. Therefore, although the results reported for halo error showed that use of the MSS produced less halo error than did the use of the BES, in reality all that the decrease in halo error on the MSS may reflect is that the responses were random. That is, as the interrater reliability results showed, the reliability coefficients reported for the differences in profile shape approach zero for the MSS. Since the estimate of reliability of the difference for shape is related to halo error, it may be inferred that the low halo error associated with the MSS resulted simply from totally unreliable ratings on this scale.

Interrater Reliability

The discussion of the findings for interrater reliability is based on the different methods used in the assessment of reliability and the implications of the results. In the present study, interpretation of the results indicated that the BES was able to differentiate among differences in the level of performance across all ratees. However, as is evident from the reliability coefficient for the MSS, there is little evidence that the MSS is reliable in differentiating among levels of profiles. Therefore, from these results we may conclude that the BES may be useful for administrative purposes such as promotional considerations and pay determination. Additionally, from this viewpoint we may conclude that the BES is a better scale type in producing higher and more meaningful reliability coefficients when examined in terms of the differences in levels of profiles.

The second test of interrater reliability provided little supporting evidence for either scale type. Neither the MSS nor the BES produced acceptable reliability coefficients. Hence, neither scale in this study was useful for providing information for developmental purposes due to the fact that neither scale type was useful in differentiating among the differences in shape (i.e., a ratee's relative strengths and weaknesses).

The results of the analysis conducted to assess interrater reliability are not very encouraging in relation to previous studies in which higher reliabilities have been obtained. Plausible reasons for this occurrence are posited and future research considerations are then discussed.

In examining the results for interrater reliability it is important to note that the ratings were conducted in a politically, volatile situation. Thus, there were a number of factors which may possibly have been of some influence on the final ratings (i.e., allegations of past race discrimination, court hearings, etc.), and perhaps may have caused the raters to be unwilling to provide careful ratings. Due to the above mentioned factors, any performance evaluation instrument may have been viewed negatively. That is, performance evaluations may have been viewed as instrumental in the perpetuation of past, discriminatory practices. Therefore, although the two scale types are both designed to create and facilitate rater involvement, it is possible that in this situation scale format--including the developmental procedures themselves (i.e., participation)--was not sufficient to counteract the fact that the raters were neither motivated or dedicated.

A second factor to consider was that the rater selection in this study was not optimal. As noted above, the raters in this study selected themselves by indicating whether or not they had sufficiently observed a given ratee's performance. Thus, the degree of familiarity that the raters had with given ratees varied. This differentiation in the familiarity that raters had with the ratee's performance reduces the reliability of the ratings.

A third and equally important reason stems from the levels of halo and leniency error associated with the two scale types. In the above, it was noted that the MSS was associated with lower levels of both leniency error and halo error. The differences noted between the two scales on these psychometric properties may have a direct relationship with the obtained levels of interrater reliability. By reducing leniency error and halo error through the use of the MSS the level of interrater reliability may also have been indirectly reduced.

Finally, as Schneier (1977) has pointed out, differences in the rater's cognitive complexity may have an impact on certain psychometric properties (i.e., halo error and leniency error). According to Schneier, "cognitive complexity is defined as the ability to discriminate between dimensions attributed to stimuli (i.e., differentiation) and the ability to discriminate within each dimension (i.e., articulation)" (p. 542). He notes that when the rating scale's complexity is matched with that of the rater the subsequent ratings are psychometrically superior. As the MSS is the more complex of the two scales (the rater must differentiate among more performance statements), it is plausible that this complexity may have contributed to the differences in interrater reliability between the two scales. As mentioned above, the obtainment of additional ratings on each dimension should lead to increased interrater reliability; however, when the complexity of the rating scale differs from that of the rater a decrease in interrater reliability may be the results.

In conclusion, it is important to note that although the two scale formats were developed and used in a manner consistent with recommended procedures, the desired scale properties did not materialize. Certain differences between the two scales may have accounted for this (i.e., the BES was not rescaled when leniency was detected, scales were not counter-balanced, etc.).

Future Research

It is important to note that although the MSS was the superior scale type in the reduction of leniency error and halo error, the MSS exhibited relatively little or no interrater reliability. For this reason, future research should consider ways to increase interrater reliability while maintaining the lower levels of leniency error and halo error. Additionally, the BES, while exhibiting higher leniency error and halo error, did exhibit relatively-high interrater reliability.

For these reasons, future research in this area should attend to four important variables. The first of which is the training of raters. Raters should be trained not only on how to eliminate certain rating scale errors, they should also be familiarized with all of the scale types which are to be utilized in a given study in order to reduce any moderating effect that scale familiarity may have.

A second important consideration is that all raters be given approximately the same opportunity to observe the ratee's performance. Thus, some type of rater rotation is suggested. By allowing raters a greater opportunity to observe the performance of a larger number of ratees, the standards used in the evaluation of performance may be more accurate. Another variable to consider is that comparable effort be exerted in the scale's development and aministration. Future research should demand equally stringent, developmental techniques for all scale types. Additionally, procedural efforts should be made to insure conformity in scale administration (i.e., counterbalance scales).

Finally, future research in this area should attend to the larger organizational context in which ratings occur in order to provide a better understanding of the variables that impinge on the performance rating behavior (i.e., environmental factors, political factors, organizational factors, rater's education, etc.).

APPENDIX A

The Behavioral Expectation Scale

LEARN AND APPLY

Learn and Apply - Firefighters must be able to understand and learn all training material including technical firefighting procedures and use of equipment, and adapt and apply this knowledge to specific situations. On this sheet, evaluate this Firefighter on the "ability to learn and apply learning" only.

Learns new material quickly and retains this knowledge. Shows considerable ability in applying knowledge gained in training to the job.	□	Upon arrival at a tank truck fire, this Firefighter could be expected to notice the label identifying the contents as highly hazardous and could be expected to remember from school where the control and shut-off valves are, and thus, shut off the valves to the truck.
	□	This Firefighter could be expected to use a Shepherd hook successfully and without error when responding to a fire in a multi story building, even though this evolution is seldom used.
Usually able to learn classroom material but sometimes has trouble adapting this learning to actual situations.	□	This Firefighter could be expected to place a tarp correctly on some personal belongings to avoid much water damage.
	□	This Firefighter could be expected to ventilate second floor windows of a warehouse from windward to leeward.
	□	This Firefighter could be expected not to understand training operations without going over them several times.
Has trouble grasping material presented in training. This becomes evident when the need arises to apply training to the job.	□	You could expect to have to show and tell this Firefighter what to do at a fire.
	□	If someone asked this Firefighter for some first-aid advise, he could be expected to incorrectly advise applying a tourniquet.

JUDGEMENT UNDER STRESS

Judgement Under Stress - In emergency situations, a Firefighter must be able to quickly, calmly, and accurately size-up situations and to make objective decisions and act accordingly. He must be able to improvise when necessary and to maintain alertness to hazardous conditions. On this sheet, evaluate this Firefighter on "judgment under stress" only.

	□	If this Firefighter entered a room as the occupant was about to leap from a window because of smoke and fire, this Firefighter could be expected to persuade the occupant to accompany him down the stairway he ascended.
Remains calm and makes effective, quick decisions under stress and pressure of emergency situations.	□	This Firefighter could be expected to jump on a hose that burst to prevent injury to others.
	□	If the side rail of a ladder broke while being climbed at a fire, this Firefighter could be expected to immediately take a pike pole and prop it under the rung of the ladder, using it as an emergency substitute for the side rail.
Usually calm in stressful situations.	□	If this Firefighter heard the cracking of ceiling beams, instead of trying to run clear of the falling roof, he could be expected to crawl under a heavy duty lathe that is nearby.
	□	This Firefighter could be expected to panic if he though the roof was falling in and possibly dive down a stairway injuring himself.
Tends to overreact or become excited when subjected to stressful situations. Sometimes makes hasty decisions when under pressure.	□	This Firefighter could be expected to become excited and want to move a victim of a serious accident without applying proper procedures.
	□	This Firefighter could be expected to get excited at a garage fire and push over a 50-gallon gas drum, thinking it was water.

TEAMWORK

Teamwork - A Firefighter must be dependable as a team member, coordinating his work with others, and carry a fair share of ALL work. A Firefighter should accept orders without questioning and maintain regular attendance. On this sheet, evaluate this Firefighter on "teamwork" only.

□ Even if this Firefighter was injured, he could be expected to stay with the man on the line until someone else showed up because he wouldn't want to leave the man on the line by himself.

Very dependable member of the team. Willing to sacrifice personal "glory" for the betterment of the team. Genuinely concerned with other team members and willing to sacrifice for them.

□ This Firefighter could be expected to notice when a fellow worker is tiring while using some tool such as an axe or pike pole and tells him to take a break and give him the tool.

□ This Firefighter could always be expected to share the work in cleaning tools and equipment and changing lines.

□ Instead of grabbing the line for himself, this Firefighter could be expected to anchor a loose ladder so that a fellow member can ascend the ladder to extinguish the fire.

Does fair share of work. Can usually be counted on to be there when needed and to help others.

□ This Firefighter could be expected to drift away from his assigned duties at a fire to perform other duties.

□ This Firefighter could be expected to be very bad with his housework and most of the time his fellow workers find it easier to do his work for him than to try to make him do it.

Sometimes trys to avoid the drudgery work. Attendance may be below that of the other members. Sometimes is not where he should be at a fire.

□ If an officer is not watching this Firefighter, he could be expected to do nothing.

PHYSICAL FITNESS

<u>Physical Fitness</u> - A Firefighter must be both physically capable and confident to perform in all stressful situations requiring physical strength and stamina. Lifting, carrying heavy objects, agility, and balance, handling equipment; stamina and endurance for extended periods are required of a Firefighter. On this sheet, evaluate this Firefighter on "physical fitness" only.

	□	This Firefighter could be expected to work for hours without a break, digging in rubble to free tornado victims.
Keeps in good physical condition. Able to respond to situations and still have enough stamina remaining to continus his normal duties at a fire.	□	After two men unsuccessfully try to pull a closet wall from this burned out portion of a building, this Firefighter could be expected to pull it off quickly by himself.
	□	This Firefighter could be expected to carry two sections of 2½" line up several stories in a burning building.
Able to respond to physically demanding situations, however, may become temporarily exhaused and not able to continue normal duties until rested.	□	This Firefighter could be expected to be unable to disconnect the hose-butts at a fire because of sore wrists.
	□	After dragging a line up to the third floor, this Firefighter could be expected to have to rest while the fire progressed.
Not concerned with physical condition and because of this, is not able to respond satisfactorily to physically demanding situations. Sometimes, is not able to maintain normal physical activity for the duration of a fire.	□	This Firefighter could be expected to be unable to carry one side of the light plant up a number of flights of stairs.
	□	This Firefighter could be expected to be unable to help raise a ladder.

COMPATIBILITY

Compatibility - A Firefighter must be able to get along and adjust to a variety of different personalities in close proximity in the Firefighter living and working situation. He must be reasonable, considerate, cooperative and have a desire to compromise and avoid conflicts. On this sheet, evaluate this Firefighter on "compatibility" only.

	□	This Firefighter could be expected to go out of his way to make substitutes feel welcome in the house.
Promotes harmony within the fire house and is sensitive to the personalities and problems of others.	□	This Firefighter could always be expected to follow the majority vote for TV programs.
	□	This Firefighter asked another who seemed to have a problem to talk about it.
Usually gets along with other Firefighters. Doesn't usually start arguments but does sometimes get involved in them.	□	This Firefighter could be expected to eat food he doesn't like just to get along.
	□	This Firefighter could be expected to ride another person about things that bother him most.
Tends to be argumentative or moody. Not usually concerned with the feelings or personalities of other Firefighters.	□	This Firefighter could be expected to tie up the cooking stove while others have to wait a long time.
	□	This Firefighter could be expected to come to work grumpy and complain about his personal problems.

PUBLIC RELATIONS

Public Relations - Firefighters must deal courteously and effectively with the public and gain their support. Taking time to help answer questions and talk to visitors indicates an interest and contributes to good public relations for the entire fire department. On this sheet, evaluate this Firefighter on "public relations" only.

Goes out of the way to project a good image of the Fire Department. Responds sincerely and courteously to questions, comments, and complaints of citizens. Shows genuine concern for citizen or victims in need.	□	In cold weather, this Firefighter could be expected to give his jacket to a person who has lost his clothing in a fire.
	□	During an inspection, this Firefighter could be expected to talk calmly to an owner who was upset with the code, explain why the code was written and thus calm the man down.
Usually maintains a professional bearing which reflects well upon the department. When interacting with the public, is usually considerate and courteous, but sometimes speaks impulsively and in some cases might be drawn into arguments.	□	While at a fire, this Firefighter might be asked by bystanders what was going on. This Firefighter could be expected to calm them down by explaining why they were breaking windows.
	□	This Firefighter might recommend various agencies to contact for help to a person who has lost his clothing in a fire.
Often not concerned with the image he projects or the image of the Fire Department. Either avoids dealing with the public by ignoring their questions or overreacts to the public by speaking impulsively. May sometimes become abusive to citizens.	□	This Firefighter could be expected to yell at a man to stay out of a fire area without letting the man explain that he lived in the building.
	□	In an incident involving a boisterous and antagonistic citizen this Firefighter might be expected to lose control and become involved in an argument.
	□	On a cold day when some children are gatherine around to watch Firefighters combat a fire, this Firefighter could be expected to turn the hose on the children.

PRIDE AND DEDICATION TO CAREER

Pride and Dedication to Career - A Firefighter must have interest and enthusiasm in the job, motivation to perform even routine duties well and willingness to support and make personal sacrafices for the rules and goals of the Fire Department. A Firefighter realizes personal shortcoming and tries to correct them. A Firefighter views the fire service as a career and maintains acceptable appearance. On this sheet, evaluate this Firefighter on "pride and dedication" only.

	□	This Firefighter could be expected to study a great deal during his free time to keep himself up on new ideas in the firefighting field.
Considerable interest in the job. Works to keep up with new materials, equipment and knowledges to increase job effectiveness. Strives to advance in the fire service.	□	This Firefighter could always be expected to encourage new Firefighters, telling them of the advantages of being a Firefighter.
	□	This Firefighter could be expected to help other Firefighters wash windows on the second floor after he has finished his section on the first floor.
Views the fire service as a career and maintains interest in the job.	□	This Firefighter could be expected to cut his hair (according to the very strictness of) regulations even though he prefers longer, more stylish hair.
	□	This Firefighter could be expected to dress sloppily and never be clean shaven.
Tends to view the job of Fire Fighter as just another job used to draw a paycheck.	□	You could expect this Firefighter not to care if he advances in the Fire Department.
	□	You can't expect this Firefighter to have any interest in keeping up the Firehouse.

OVERALL PERFORMANCE RATING

Now that you have rated your employee on the seven rating scales, your next task is to give the employee an overall rating on his total job performance. Please use the following scale by circling the appropriate value.

Very Low			Average			Very High
1	2	3	4	5	6	7

Appendix B

The Mixed Standard Scale

FIREFIGHTER

EXPERIMENTAL PERFORMANCE RATING

FORM MS

Rater ____________________________

Ratee ____________________________

Please read and consider the following statements one at a time. You are asked to circle "+" if the ratee is better than the statement. "O" if the statement fits the ratee, "-" if the ratee is worse than the statement. Circle only one response to each statement.

+ = the ratee is better than the statement

O = the statement fits the ratee

- = the ratee is worse than the statement

+ O - As this Firefighter heard the cracking of ceiling beams, instead of trying to run clear of the falling roof, he crawled under a heavy duty lathe that was nearby.

+ O - In cold weather, this Firefighter gave his jacket to a person who had lost his clothing in a fire.

+ O - This Firefighter carried two sections of 2½" line up several stories in a burning building.

+ O - This Firefighter asked another who seemed to have a problem to talk about it.

+ O - This Firefighter goes out of his way to make substitutes feel welcome in the house.

+ O - After dragging a line up to the third floor, this Firefighter had to rest while the fire progressed.

+ O - This Firefighter helped other Firefighters wash windows on the second floor after he had finished his section on the first floor.

+ 0 - This Firefighter dresses sloppily and is never clean shaven.

+ 0 - This Firefighter rides another person about things that bother him most.

+ 0 - This Firefighter panicked because he thought the roof was falling in and dove down a stairway injuring himself.

+ 0 - This Firefighter does not understand training operations without going over them several times.

+ 0 - This Firefighter yelled at a man to stay out of a fire area without letting the man explain that he lived in the building.

+ 0 - This Firefighter placed a tarp correctly on some personal belongings to avoid much water damage.

+ 0 - While at a fire, this Firefighter was asked by bystanders what was going on. This Firefighter calmed them down by explaining why they were breaking windows.

+ 0 - This Firefighter shares the work in cleaning tools and equipment and changing lines.

+ 0 - This Firefighter studies a great deal during his free time to keep himself up on new ideas in the firefighting field.

+ 0 - This Firefighter jumped on a hose that burst to prevent injury to others.

+ 0 - Upon arrival at a tank truck fire, this Firefighter noticed the label identifying the contents as highly hazardous and remembering from school where the control and shut-off valves are shut off the valves to the truck.

+ 0 - This Firefighter drifts away from assigned duties at a fire to perform other duties.

+ 0 - This Firefighter worked for hours without a break, digging in rubble to free tornado victims.

+ 0 - Even though this Firefighter was injured, he stayed with the man on the line until someone else showed up because he didn't want to leave the man on the line by himself.

Appendix C

Scoring Combinations for the Mixed Standard Scale

Scoring Combinations for Mixed Standard Scale

Logical Response Scoring

Statements			Points
I	II	III	
+	+	+	7
0	+	+	6
-	+	+	5
-	0	+	4
-	-	+	3
-	-	0	2
-	-	-	1

Illogical Response Scoring

Combination			Points
I	II	III	
+	+	0	7
+	+	-	7
0	+	0	6
0	+	-	6
-	+	0	5
-	+	-	5
0	-	+	5
0	0	-	4
+	0	+	4
+	0	-	4
0	0	0	4
-	0	-	3
+	-	+	3
+	0	0	3
0	-	0	2
+	-	0	2
+	-	-	1
0	-	-	1

BIBLIOGRAPHY

BIBLIOGRAPHY

Arvey, R.D. & Hoyle, J.C. A Guttman approach to the development of behaviorally based rating scales for systems analysts & programmer/analysts. *Journal of Applied Psychology*, 1974, *59*, 61-68.

Bernardin, H.J. Behavioral expectation scales versus summated scales: a fairer comparison. *Journal of Applied Psychology*, 1977, *82*, 422-427.

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

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

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.

Blanz, F. & Ghiselli, E.E. The mixed standard scale: a new rating system. *Personnel Psychology*, 1972, *25*, 185-199.

Blood, M.R. Spin-offs from behavioral expectation scale procedures. *Journal of Applied Psychology*, 1974, *59*, 513-515.

Blum, J.L. & Naylor, J.C. *Industrial Psychology*: *its theoretical and social foundations*. New York: Harper & Roy, publishers, 1968.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 1975, *60*, 550-560.

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

Borman, W. C. & 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*, 1974, *59*, 197-201.

Burnaska, R.F. & Hollmann, T.D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology*, 1974, *59*, 307-312.

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

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

Finley, D.M. & Osburn, H.G. & Dubin, J.A. & Jeanneret, P.R. Behaviorally based rating scales: effects of specific anchors & disguised scale continua. Personnel Psychology, 1977, 30, 659-669.

Friedman, B.A. & Cornelius, E.T. Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 1976, 61, 210-216.

Keaveny, T.J. & McGann, A.F. A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 1975, 60, 695-703.

Latham, G.P. & Wexley, K.N. & Pursell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

McPahil, S.M. & Dickinson, T.L. MSS: A program for scoring mixed standard scales. Applied Psychological Measurement, 1977, 1, 402.

Snedecor, G.W. & Cochran, W.G. Statistical Methods (6th ed.). Iowa: Iowa State University Press, 1967.

Saal, F.E. & Landy, F.J. The mixed standard rating scale: An evaluation. Organizational Behavior and Human Performance, 1977, 18, 19-35.

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.

Nunnally, Jum C. Psychometric Theory. St. Louis McGraw-Hill Book Company, 1967.

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