THE EFFECTS OF SCALE CONTINUITY AND BEHAVIORAL ANCHOR SPECIFICITY UPON THE PSYCHOMETRIC PROPERTIES OF PERFORMANCE RATING SCALES

A Dissertation

Presented to

the Faculty of the Department of Psychology

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

David M. Finley

December, 1975

ACKNOWLEDGMENTS

While the list of contributers to this research is long, I still wish to acknowledge my gratitude to all those individuals most directly involved. Without exception, their efforts were significant.

First, my sincere thanks to my major professor, Dr. H. G. Osburn, whose selfless giving of both time and energy went far beyond any professional obligation. The suggestions and support of the other committee members, Drs. James Campion, Stuart North, and Daniel Sheer were also greatly appreciated. Bob Paver and Dr. Ira Chorush, by providing their expertise in programming and statistics respectively, considerably reduced the time and hassle required to complete the study.

Two dear ladies of similar name and nature were as supportive with their kindness as they were with the clerical skills which they cheerfully supplied. Ms. Betty Bowers typed much of the first draft, while Ms. Betty Baldwin typed the final copy. Ms. Vicki V. Vandaveer spent countless hours collating and mailing the myriad assortments of rating scales, without once protesting the absurdity of such a complex scheme.

Acknowledgments typically conclude with recognition of the long-suffering spouse, whose moral support made the research burden bearable. The contributions of Janet Havis (Finley) included but went far beyond mere moral support. Her inputs were genuine technical and scientific efforts for which she is endeared even more. To paraphrase a song, she's simply the best thing that ever happened to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (Cont.)

LIST OF TABLES (Cont.)

LIST OF FIGURES

CHAPTER I

INTRODUCTION

For the past three decades a recurring theme has characterized the introductions of most performance evaluation literature: "the task of developing reliable and valid job criteria is one of the most challenging and desired objectives in contemporary organizational psychology." Admittedly, numerous technical refinements have occurred since the fundamental formulations of the "criterion problem" were presented by Otis (1940), Bellows (1941), Toops (1944) and Thorndike (1949). Still, the controversies concerning criteria problems continue, perhaps with accelerating vigor.

Lawler (1967) notes that it has become increasingly stylish to criticize the most frequently used measure of job performance--the superior's subjective rating of the subordinate's work effectiveness. Yet, the superior's evaluation continues to be employed more often than any other performance measure. Whether used for research purposes, such as test validation and training evaluation, or for personnel records maintained for decision-making purposes, subjective measures of job performance outnumber the more objective measures at least two to one (Vroom, 1964; Guion, 1965). Even in the face of increased pressures from the Equal Employment Opportunity Commission and the Office of Federal Contract Compliance, researchers have continued to rely primarily on subjective supervisory evaluations (Bray & Moses, 1972).

This is not to say, however, that there has not been an extensive effort to find objective criteria, free from the contaminating effects characteristic of subjective evaluations. Direct measures have been attempted even at the managerial level, where performance is most difficult to quantify and make objective. Bingham and Davis (1924), Guifford (1928), Hulin (1962) and Williams and Harrel (1964) have used salary as a criterion of job performance. Henry (1948), and Starch (1942) selected organizational level achieved as a measure of performance. Guion (1965) reports the use of objective criteria in studies by Comrey, High and Wilson (1955); Robbins and King (1961); Cuomo (1955); and MacKinney and Wolins (1960). None of these have shown promise, however, for measuring managerial performance.

The impracticality of many direct measures of job proficiency is that, first, considerable time and effort is often required for data collection; secondly, these indicators are in most cases more seriously contaminated than are superior's evaluations. Salary level and promotions, for example, are based upon subjective decisions and in addition are at least partially determined by factors both irrelevant to job performance and outside the control of the individual being evaluated.

Faced with a long history of unsuccessful efforts to adequately develop either objective or subjective performance criteria, researchers have directed their attention toward the reasons why evaluation procedures typically fail.

Ghiselli (1956) was among the first to point out that "far more attention has been devoted to the development of predictive devices than to the understanding and evaluation of criteria." (p. 1) Campbell, Dunnette, Lawler, and Weick (1970) contend that:

> Psychologists have not paid proper heed to the extreme difficulties involved in accurately observing and reporting behavior. They have given insufficient attention to meaningfulness, behavior definition, and semantic clarity in their development of job behavior rating scales, with the result that most scales either have not been understood or were viewed as irrelevant by observers asked to complete them.

These authors therefore argue, not for the abolition of subjective evaluations, but argue instead for the development of methods to avoid the mistakes which they have described.

Research concerning merit rating has focused mainly on two areas: the reliability of the rating and the reduction of rating errors; the two are related, of course. No generally valid solution of the reliability problem has yet been found. Likewise, it has not yet been proved possible to reduce the sources of error to an acceptable level. In fact, most of the sources of error seem to be such that their total elimination is impossible. Nevertheless, some recent innovations in the evaluation of work effectiveness have shown considerable promise. It is the purpose of the present study to experimentally contrast the two techniques which are currently receiving considerable research attention. The proponents of these two methods seem to make similar claims regarding the effectiveness of their technique in reducing

evaluation error and increasing reliability.

First, a review of the performance evaluation literature is presented including: a categorization of evaluation methods, discussion concerning the sources of unreliability and error associated with each category, and factors independent of the method of evaluation which affect the validity of performance measures.

CHAPTER II

RELEVANT RESEARCH: A REVIEW OF THE LITERATURE

Classification of Evaluation Methods

The different methods of evaluating job performance have been classified according to a number of different schemes. Since a wide variety of modifications and combinations of evaluation methods have been described in the literature, no classification system can be totally unambiguous.

Ghiselli (Ghiselli & Brown, 1955) divided evaluation methods into four classes:

1. Ranking methods -- those concerned with rank order but taking no account of the size of differences between individuals.

2. Rating-scale methods -- those using units of measurement and scales for the measurement of the magnitude of the differences between individuals.

3. Check-list methods -- those where the evaluator is provided with a list of scaled definitions, from which he selects those best describing the individual.

4. Forced-choice methods -- those in which the evaluator is forced to make a choice between two or more different descriptions of behavior which are equally weighted in terms of job success.

Guilford (1954) presented a classification system based upon the technique of measurement applied: numerical,

graphic, standardized, forced-choice and cumulative points methods. This system, though relatively rough when presented alone, could be used to further subdivide the categories presented by Ghiselli.

Tiffin (1959) chose to divide evaluation methods into six categories which overlap considerably with Ghiselli's scheme: graphic rating scales, ranking methods, paired comparisons, forced-choice methods, forced distribution methods, and critical incidents methods. Tiffin's classification system seems somewhat confounded. For example, forced distribution can be incorporated in almost any method. Similarly, the critical incident method frequently employs the technique of the checklist method.

Anastasi (1964) proposed seven categories of evaluations: order-of-merit comparisons, rating scales, scaled check-lists, the forced-choice technique, critical incidents, the nominating technique, and the field review method. Only the last two methods have not been included in previously mentioned classifications. The nominating technique is a method of peer rating while the field review method is based upon information gained through interviews with the individual's superiors.

Blantz (1965) recognized a classification based upon the relationship between the evaluator and the individual being evaluated. This system distinguishes evaluations by superiors, peer ratings, subordinate ratings, and self-evaluation methods.

Guion (1965) presents a straightforward system using three main classes:

1. Rating scales -- including any format in which the rater is presented with a visual scale.

2. Employee comparisons -- where one employee is compared with another or a group of others. This category includes the methods of rank order, paired comparison, forced distribution, and comparison between groups.

3. Checklists -- including the technique of summated ratings, equal-appearing intervals, and forced-choice ratings.

Smith (in press) classified evaluations into five categories according to rating scale format:

1. Direct estimation formats -- those that ask directly the question, how good is the ratee? The answer is typically recorded on a vertical or horizontal scale. Numerical, alphabetic, descriptive statement and symbolic anchors all have been utilized to identify performance level along the scale.

2. Ranking formats -- those that order individuals without asking for direct estimations of distance along a scale. Ranking individuals from high to low on any relevant dimension, paired comparisons of every individual with every other individual, and ordering through the use of forced distributions according to fixed percentage categories are all included in this method.

3. Test construction formats -- methods that can be used to achieve a scale of cumulative points. Included here

are: unweighted checklists in which adjectives or statements of critical incidents are checked as applying or not applying to the ratee; and weighted checklists which include scaled behavioral expectations and semantic differential scales.

4. Items as scaled standards -- formats that set up a series of items as scaled standards against which an individual may be judged. These items may be used as anchors, usually along a graphic rating scale or in forced-choice format.

5. Individuals as scaled standards -- a method using actual people as anchors against which employees can be compared. Each rater evaluates not only his own subordinates, but also appropriate reference persons in other departments. The evaluations of reference individuals are then transformed into standard scores in order to (1) permit cross-referencing of different raters with different groups of subordinates and (2) take into account the anchoring of individual rating scales.

Smith's classification system has much in common with those of Ghiselli, Guion, and Tiffin. Like that of Tiffin, however, her system is a bit ambiguous. For example, it would seem that scaled behavior expectations presented in graphic rating scale format could in many cases be assigned to both her first and last categories.

Of the various classification systems discussed, the highly similar categories of Ghiselli and Guion, while not

totally comprehensive, seem the most unambiguous and straightforward. Because it is based upon the structure of the rating form and upon the method of response, Ghiselli's method of classifying evaluations appears the most appropriate for the purposes of the present study. His system has therefore been utilized in organizing this review of the merit rating literature.

Ranking Methods

Using this method, persons to be evaluated can be rank-ordered in comparison to other individuals or groups of others. Comparison can pertain to a single, global dimension or to several independent traits. There are essentially four methods of establishing a rank-order.

1. Complete ranking -- where individuals are simply ordered from best to poorest along a relevant dimension. Some administrators permit ties in cases where discriminations cannot be made, while others specify that no two persons may occupy the same rank. One recommended administrative procedure is alternation ranking. The evaluator alternately selects his best and then his poorest employee on the dimension in question. This procedure is repeated with the remaining employees until all have been ranked.

Ranking is a relatively simple administrative procedure. Arguments have been made for its use where raters are unsophisticated and it is not feasible to provide them with evaluation training.

Among the more negative aspects of complete ranking is the fact that the rater's job becomes quite laborious with larger groups. Also, discriminations become progressively difficult near the center of the distribution. Inter-individual differences are likely to be partly artificial and fortuitous, particularly within the middle range.

Ranking provides only an ordinal scale which does not indicate the amount of difference between people. However, ranks may be treated under a number of assumptions (Guilford, 1954) in order to provide estimations of distances between people and to permit the combination of rankings of different persons by more than one rater.

2. Forced distribution -- where individuals are assigned to order-of-merit categories. This is a gross ranking in that a rater is not required to distinguish between workers who are approximately equal on the dimension in question. Therefore, the rater's job is easier and less time consuming.

This method is a variation of the graphic rating scale designed to control leniency and central tendency errors. A prescribed percentage of employees are placed in each category, forming essentially a normal distribution. The assumption of a normal distribution is seldom warranted in organizational populations which have been preselected, discharged, promoted, etc. Also, the method has been shown highly susceptible to rater bias (Klores, 1966).

3. Paired comparisons -- where every employee is compared with every other employee. The advantage of this method is that the rater need not keep the entire group of employees constantly in mind. Instead, he is asked to concentrate on the comparison of only two persons at a time.

The paired comparison method has the disadvantages of (1) requiring the rater to make a large number of judgments even when evaluating a small group of people, and (2) considerable data computation. Guilford (1954) suggests that paired comparison ranking is too cumbersome for evaluating more than fifteen people. Lawshe, Kephart, and McCormick (1949) disagree. They developed a system by which twenty-four names may be rated reliably in thirty minutes. Others (McCormick and Bachus, 1952; Guilford, 1954; Forgerson, 1958) have presented methods to reduce the number of paired judgments required. McCormick and Bachus found, however, that reliability diminished almost systematically as a function of the pairs omitted.

4. Comparison between groups -- where an individual's ranking is compared in relation to other members of the organization in addition to his own evaluation group. This method was previously discussed in an illustration of Smith's (in press) classification category "individuals as scaled standards." Uhrbrock and Richardson (1933) and Rosensteel (1953) used this method where several key men, known to all raters, could serve as anchors against which employees were compared. The average evaluations of the key men can be

taken as reference points on a scale. The ranking of all other employees, converted to normalized scale values, can then be plotted in relation to the reference points on the scale.

Ideally, this method requires that the same key men be used by all evaluators and that the performance of these men be equally familiar to all evaluators. Ross (1966) proposed a solution to this problem. Each rater includes in the ranking of his subordinates any individuals outside his department whose performance he can assess and who are comparable to benchmark persons in his own department. The evaluations of all benchmark individuals are then transformed into a standardized score based upon their ranks within the reference groups and their standardized score within their own department. This procedure, while amenable to a global evaluation, seems too involved for use in evaluating individuals on several dimensions.

Psychometrically, the most serious limitation of the various ranking methods is their restriction to ordinal measurement. Using rankings it is impossible to detect the presence of a skewed distribution. From the applied standpoint, the information that ranking methods offer is frequently not sufficiently differentiated. They seem better suited to global estimates of job performance (although individuals can be ranked on two or more performance dimensions) and are therefore restricted in their use. Such overall estimates tend to reflect common stereotypes of

successful performance rather than affording diagnostic information about relevant job performance.

Rating Scale Methods

All evaluation methods which permit the creation of scales might be included in this group. In that case, the check list and forced-choice methods could also be regarded as special forms of rating scales. However, in this review the term is used in a narrower sense. Only those methods in which the scale is directly at the rater's disposal are included. Based upon the structure of the form and the method of response, four types of rating scales are discussed.

1. Numerical scales -- where the rater is presented a sequence of numbers, each representing a certain degree of the trait in question. Typically, some type of adjective, phrase, or definition is affixed to at least a few of the numerical values along the scale.

One of the strongest arguments for the use of numerical scales is their ease of construction. Also, the evaluations obtained can be treated as interval measurements. The weak points of numerical scales will be covered in a discussion of the types of rating errors associated with evaluations in general.

2. Graphic scales -- where the rater is presented a segmented or continuous straight line, along which he locates the individual being rated. Scale values are obtained by measuring the distance of each mark from the origin. Where the graphic scale forms a continuum, the

rater is not confined to a limited number of categories as in the case with numerical scales.

Graphic rating scales are also simple and easy to construct. Furthermore, discrimination between the persons to be rated can be treated as minutely as desired.

3. Descriptive adjective scales -- where an attempt is made to give the rater a concrete picture of each category along the scale through the use of descriptive adjectives such as "good," "satisfactory," "fair," "poor," etc. Combinations of descriptive adjective and numerical scales are encountered frequently in practice.

Numerical, graphic, and adjective rating scales enjoy wide use in merit rating. However, they are all subject to serious sources of error which often mar the reliability of the ratings and therefore reduce the usefulness of the evaluations. These sources of error will be examined at length.

4. Behaviorally based scales -- where the scale values of the trait concerned are identified by descriptions of actual job behavior. First developed by Hartshorne and May (1929) the critical component of this method rests on the characteristics of the behavioral definitions. The definitions of a particular trait must be carefully chosen, unambiguous, concrete, and highly characteristic of that particular dimension. To achieve these objectives requires careful job analysis and the assistance of individuals familiar with all aspects of the target job positions.

Obviously, a behaviorally based rating scale which meets the above requirements is far more difficult and time consuming to construct than the previous forms of rating scales discussed. However, the practical value of a rating scale depends heavily upon the properties of the definitions contained within it. Yuzak (1961) and Stockford and Bissell (1967) found "a marked influence on the value of ratings" in favor of descriptive scales over evaluative scales.

The advantage of the behaviorally based scales appears to be in their ability to keep the evaluator "on track" in terms of distinguishing between different traits and between the relative degrees of any one trait. Where a trait is defined only by a title or short evaluative phrase, different raters may attach different interpretations to the dimension. The result is that two raters, faced with the same evaluation stimulus, may rate individuals on different performance factors. In addition, lacking information as to the inclusiveness of the particular dimension, the rater typically expands his frame of reference to include characteristics of the ratee which belong to another evaluation dimension.

In contrast to specific descriptions of behavior, the various numerical, graphic, or adjective reference points employed within a scale remain somewhat abstract. For example, numbers do not in themselves bear any concrete relationship with actual behavior. Therefore, the rater must mentally create more concrete counterparts to the numbers. Adjectives may also be understood differently by

different evaluators. As an example, raters tend to disagree which adjective "superior" or "excellent," is of higher order. An experiment by Ghiselli (Ghiselli & Brown, 1955) provides an illustrative example. Subjects arranged fourteen adjectives describing job success in order from best to worst. There were large differences of opinion concerning the hierarchy of some adjectives. The adjective "satisfactory," for instance, varied from fifth to eleventh place.

To the extent that the behavioral descriptions comprising a scale are unambiguously stated, observably concrete, and unidimensional, ratings by different people are better comparable with one another. When these requirements are met, it is far more probable that different raters will start from the same reference point, evaluate the same dimension, and utilize the same interval scale. Oksala (1958) strived to meet these difficult requirements in developing scales to measure the job success of workers in a Finnish metal product factory. Factor analysis of the ratings obtained a number of different factors, indicating that different traits were distinguishable as opposed to a global evaluation.

Check-list Methods

These methods usually share three characteristics: first the rater is reporting rather than evaluating what the ratee does. Secondly, in filling out the check-list,

the rater is generally not aware of the order-of-merit class to which he is assigning the ratee. Thirdly, the rater is usually free to vary the number and nature of the traits checked.

A typical example of the check-list method consists of a large number of statements or adjectives, usually of both positive and negative connotation. The rater is required to check all statements or adjectives which apply to the ratee. If the rater is unaware of the scoring value of each statement or adjective, he cannot know with certainty how good or bad the ratee's final score will be.

The first to employ the method was Thurstone (1928), who used it for attitude measurement. Hartshorne and May (1929) used the same technique for evaluating personality traits in children.

An early industrial application of the method was that of Richardson and Kuder (1933). They asked two raters to each evaluate salesmen for job success and found that inter-rater reliability was surprisingly high. Goertzel (Ghiselli & Brown, 1955) obtained similar results with subjects representing a variety of occupations.

A more sophisticated technique which can also be subsumed under the check-list method is that of critical incidents (Flanagan, 1954; Buel, 1960). The technique requires job knowledgeable individuals to provide observable incidents of particularly effective or ineffective performance exhibited

by job incumbents. Selected incidents can then be assembled into a check-list. The method was first developed for the military during World War II but has also been applied in industry (Kirchner and Dunnette, 1957; Stoltz, 1958; and Tiffin, 1959).

Although usually yielding an overall evaluation, the advantage of this method is its emphasis upon behavioral observation. The rating is based upon behaviors that are actually under observation and that the ratee exhibits or fails to exhibit. The critical incidents technique can be combined with other evaluation methods to make the contents more concrete for the evaluator. Kay (1959) combined it with the forced-choice method.

A major disadvantage of the critical incidents technique is the problem of collecting incidents. Smith (in press) reports that, with the exception of the military situation, "it has become necessary to use broad generalizations, losing the huge advantage of emphasis on observation." Other disadvantages include: the mechanical problems of recording and classifying incidents, the frequency of recording incidents, and the format in which they are presented. Suggested solutions to these problems have been made (Flanagan and Burns, 1955), however, there is little research evidence supporting these recommendations (Korman, 1971).

Forced-choice Methods

To an even greater extent than check-list methods, forced-choice rating attempts to counteract the potential

bias resulting from the rater's awareness of the order-of-merit class to which the ratee will be assigned. In order to conceal from the rater the meaning of his ratings, scales are constructed in such a way that it is not obvious how different responses will be scored.

The typical forced-choice format consists of groups of four statements. The rater is asked to evaluate the individual in regard to each set of statements. A set may be comprised of statements which all seem favorable, statements which all appear unfavorable, or a combination of two favorable and two unfavorable statements. Requirements for the rater can also vary. For example, the rater can be directed to choose from each set of statements the one statement most descriptive of the ratee, the one statement least descriptive of the ratee, or both the statement most applicable and the one least applicable to the ratee.

Statements in a set are matched for general desirability but differentiated in terms of their predictive ability against a criterion. The intent of this technique is to increase rater objectivity while reducing rater carelessness and deceit. If a rater is inattentive or attempting to give only favorable ratings, he is as likely to select a glittering generality which is equally applicable to effective and ineffective performance as he is to select a statement describing a genuinely important characteristic of excellent performance.

Berkshire and Highland (1953) conducted a study to

determine the item group structure most conducive to high validity and reliability. A form consisting of sets of four favorable statements, from which the two most descriptive of the ratee had to be chosen, was found superior. Regarding the number of sets of items required, Taylor, Schneider and Clary (1954) obtained almost identical results when a form containing ten sets of items was compared with a 28 set form. Lepkowski (1963) obtained encouraging results using the forced-choice method to evaluate the job success of engineers. His form required the rater to choose, for each of 20 item-triplets, the most and least descriptive statements. The split-half reliability of the ratings was +.90 and they correlated +.74 with another evaluation method using four-point rating scales.

The forced-choice method is based upon the notion that the rater should be deceived in order to improve evaluations. A rater who wants to rate all employees high can beat the system, however, by using an outstanding employee as a frame of reference for responding to the item sets. Tiffin (1959) lists three additional criticisms of forced-choice evaluations: (1) forced-choice item sets are laborious to construct; (2) raters tend to react unfavorably to the forced response, especially for wholly negative item sets; and (3) it is difficult and often impossible to keep the scoring key secret from the raters.

Common Evaluation Errors

The subjective nature of ratings results in several

sources of error which reduce both the reliability and the validity of the evaluations. These errors are shared to greater or lesser extent by all the evaluation methods requiring rater judgments. Most important are: the leniency-severity errors (the tendency toward skewed distributions), the error of central tendency (lack of dispersion) and the errors resulting from the so-called halo effects (inter-correlational errors).

Leniency and Severity Errors

Raters are generally inclined to place most ratees near one end of the scale. As a rule, the majority of individuals are evaluated as excellent or far above average employees. The result is displacement of the mean and skewness of the distribution of ratings. There may be cases where a negatively skewed distribution accurately depicts the actual population. For example, a work unit may possess only highly experienced and proficient workers because those with inferior skills and abilities have been eliminated through the process of attrition. In such a case, the rater's favorable evaluations may not be in error. Cases are often encountered, however, where the mean of the ratings is so high that it cannot be assumed to reflect the average employee's proficiency.

There are numerous reasons for highly favorable ratings (Thorndike, 1949; Bass, 1956; Sharon and Bartlett, 1969):

1. The rater may be in the position of judging his own competency along with that of his subordinates. If they are

not performing adequately, he is not a competent supervisor.

2. The rater may believe that anyone who has survived in his organization has had to perform at an above average level.

3. Ratings may be a reflection of the rater's heightened self-esteem.

4. The rater may fear the discovery of a poor evaluation by the ratee or the face-to-face performance review in which a negative evaluation would be discussed.

5. The rater may intentionally overrate subordinates in order to protect or further their interests. Taylor and Wherry (1951) found that ratings were higher when carried out for recruitment and placement purposes than when collected for research purposes only.

Merit ratings may also yield a positively skewed distribution, although this occurs far less frequently. A positively skewed distribution may be due to:

1. The rater's overcompensating efforts to avoid erring in the favorable direction.

2. An insecurity on the rater's part concerning standing in the organization and a fear that subordinates will pass him by.

3. A contrast-effect may exist in that if the rater considers himself high on a trait, he may set himself as a standard against which he sees all subordinates in a less favorable light (Murray, 1949; Johnson & Vidulich, 1956).

The four evaluation methods: ranking, rating scale,

check-list, and forced-choice vary in their approach and in their effectiveness in reducing the leniency-severity problem. The ranking methods eliminate skewness by ordering individuals along a continuum or by forcing an essentially normal distribution. Perhaps the greatest criticism of the ranking methods in this regard is that they potentially create normal distributions where skewness may in fact exist.

Check-list and forced-choice methods take another approach, attempting to prevent the rater's adjustment of ratings in a favorable or unfavorable direction by concealing from him the scoring of his responses. In the check-list methods the rater is not provided with the values of the various phrases or adjectives presented. Forced-choice statement sets provide "distractors" which are intended to appear equally favorable or unfavorable to the statements in the set which are predictive of job performance. The critical factor influencing the ability of these two methods to reduce leniency-severity error is therefore the prevention of the rater's awareness of the interpretation of his response. This assumes that the rater can neither "see through" the items nor obtain the evaluation scoring scheme--two questionable assumptions.

It is with the rating scale methods that the greatest potential for leniency-severity error occurs. With rating scales the rater is neither forced to spread his ratings, nor are the values of his ratings hidden. The rater is free to make evaluations of each and every ratee as he so desires.

Attempts to overcome the leniency-severity problems of rating scales have taken two forms: training of raters and attention to the descriptions anchoring various points on the scales.

When training is provided concerning (1) the purpose and uses of the evaluations, (2) the content of the scales, and (3) the tendency toward leniency-severity errors; rater motivation and performance should theoretically improve. Research concerning the effectiveness of rater training will be discussed at a later point in connection with factors affecting the reliability of ratings.

A second method of reducing rating scale leniency has been to attend to the statements or adjectives describing the various scale values. In particular, unfavorable statements can be modified in a more positive direction to render the distribution of rating more normal (Guilford, 1954; Blantz and Ghiselli, 1972). Campbell, Dunnette, Arvey and Hellervik (1973) advocate the use of scaled expectations to reduce leniency-severity. They contend that scaled expectations yield less leniency errors because they specifically define anchors for the performance continua and increase rater motivation by involving the rater in development of the scales.

Central Tendency Error

A second type of distribution error is committed by evaluators who tend to avoid both high and low ratings, instead rating most individuals as average and thereby

restricting variability around the center of the scale. There are several potential reasons for the error of central tendency. A rater typically hesitates to make extreme evaluations in those instances where he is unfamiliar with the performance of the ratees (Guilford, 1954). The evaluator who is insufficiently motivated can be rid of the rating task and make the fewest rating errors by placing all ratees in the middle range. The rater who wishes (for whatever reason) to avoid the proverbial limb can employ the same strategy.

Measures similar to those used to avoid skewed distributions can be employed to counteract the error of central tendency. Ranking methods force evaluation variability. Check-list and forced-choice instruments attempt to shift the evaluator's attention away from the merit value to which the ratee will be assigned. Rating scales may be constructed using behavioral anchors which the evaluators have submitted as representative of extremely effective and ineffective as well as moderate levels of performance. Careful training of raters, however, is probably the most promising deterent to central tendency error. The central tendency error is particularly characteristic of inexperienced raters (Ekman, in Blanz, 1965).

Halo Effect

Perhaps the most common evaluation error (Guion, 1965) is the tendency to allow a general, overall impression of the ratee to influence the rater's evaluation of several

distinct characteristics. This error is essentially a failure on the rater's part to differentiate between several traits being evaluated. It results in high intercorrelations among the ratings of behaviors or characteristics which may in fact be relatively independent of one another. Halo can effect ratings in either a positive or negative direction, merely implying that the rater's prevailing impression of the ratee (whether favorable or unfavorable) is the determinant of the similar ratings on each characteristic.

Halo error stems in part from the tendency of raters to value some characteristics of workers more than others. If a ratee is known to be high on one characteristic, and this characteristic is esteemed by the rater, the rater is inclined to expect that worker to be high on other traits as well. Thus, factors causing halo error are therefore dependent upon the values of the rater and may vary from rater to rater; and for one rater, may vary from ratee to ratee.

Failure on the part of raters to discriminate between the various work characteristics being evaluated can render ratings useless except as a general, overall measure of work performance. Such a global measure is especially harmful to validation studies which utilize a battery of tests, each selected to measure different aspects of job performance. Where halo in the ratings results in a simple performance factor, it is necessary to compare subjects' scores on each

subtest with this overall criterion. In such cases the ability of individual tests to predict a particular performance factor for which they were chosen is lost. Typically only those tests which correlate significantly with the general, overall criterion are retained in a final test battery.

The extent of the halo effect can be assessed through factor analysis. The greater the interference of halo, the higher the intercorrelations of the different trait ratings and the fewer factors which can be extracted. Therefore, a goal in the development of merit ratings has been to devise methods which yield several clearly distinguishable performance factors. A review of the research literature reveals that this goal has been difficult to attain by any particular method.

Many studies report a failure to obtain more than a single, general performance factor (Jurgensen, 1950a; Grant, 1955; Guion, 1961). Investigating methods to counteract halo effects Taylor and Wherry (1951) obtained better results with forced-choice than with the graphic method. Seashore and Tiffin (1952), however, were only able to extract a single factor from data using the forced-choice technique.

The techniques which prevent the rater from knowing the result of his rating may be superior to the more obvious rating techniques in reducing halo. Ewart, Seashore, and Tiffin (1941) and Sisson (1948) found the halo effect to be smaller when ratings are not obviously made on a scale.

Willingham (1958) found that in obvious scales, the rating given to a person on one trait affects the rating given him on the next trait. One high rating tends to be followed by another and vice versa. This tendency increases with the number of merit degrees presented for each trait. Therefore, in cases where the rater is able to see the result of each rating, some researchers (Guilford, 1954; Johnson and Vidulich, 1956; Wells and Smith, 1960) have suggested that halo can be counteracted by requiring the rater to evaluate one trait at a time for all ratees. This strategy diminishes the likelihood that the evaluator will remember an individual's rating for a previous trait and allow this rating to affect subsequent ratings. Johnson (1963) found, however, that this procedure is not always effective in reducing halo. He asked one group of raters to evaluate five persons each day on one trait while another group evaluated one person per day on five traits. There were no statistically significant differences between the halo effects in the ratings of the two groups.

Another approach to the reduction of halo effect is the development of groups of rating scales whose intercorrelations within a group are expected to be high, thereby comprising a factor. However, between groups the scales are intended to be relatively independent. It is critical that scale clusters have meaning for the raters. In actual studies where rating scales were developed solely on the basis of face validity--that is, scales which the researchers

judged likely to comprise independent factors--the results have not been promising. Jurgensen (1950b) was unsuccessful in obtaining independent factors by this means using either numerical scales or descriptive statement scales.

An alternative method is to have the raters develop scales which they perceive as independent factors. This is essentially the procedure used in a currently esteemed method, scaled expectations of behavior. Scaled expectations were first proposed by Smith and Kendall (1963) and have since been used by Arvey and Hoyle (1974), Borman and Vallon (1974), Burnaska and Hollmann (1974), Campion, Greener, and Wernli (1973), Folgi, Hulin and Blood (1971), Landy and Guion (1970), and Zedeck and Baker (1971), among others. This method charges appropriate organizational personnel, rather than the researcher, with the task of developing the various dimensions of job performance and defining specific behavioral anchors for each performance continua. The re-translation step, which is central to the Smith and Kendall (1963) procedure, then requires the raters to sort each behavioral incident into the dimension which they judge it represents. Only those scales containing behavior which the raters clearly distinguished from all other performance dimensions are retained.

Research with rating scales developed according to the technique of scaled expectations has typically evaluated the scales in terms of convergent and discriminant validity (Campbell and Fiske, 1959). When ratings on nine dimensions

developed by scaled expectation techniques were instead submitted to factor analysis, Campbell, Dunnette, Arvey, and Hellervik (1973) report that this procedure yielded several nontrivial factors with one relatively high loading per factor. Furthermore, a much clearer factor solution was obtained from the ratings using the scaled expectation technique than from summated Likert-type scales used for comparison purposes in the study.

Studies have been reported in which the scaled expectation technique was not employed, yet, a number of factors have emerged. However, the rating forms used in these studies have often included such a large number of traits that it would be difficult to apply these forms in actual practice. For example, Hausman and Strupp (1955) employed a form containing 55 traits and were able to obtain five factors whose intercorrelation ranged from .49 to .79. Stoltz (1959) used a rating technique which employed 250 questions. Here also, analysis of the data yielded five factors. Turner (1960) was able to obtain four factors by using objective criteria (grievances, turnover, absences, tool costs, etc.) in addition to merit ratings.

Still another device in attempting to eliminate the halo effect is careful training of the raters. Just as with the rating errors of leniency and central tendency, it would seem that the rater who is aware of the potential for halo error is better equipped to avoid this error in his ratings.

Reliability of Merit Ratings

Reliability can be defined as the "extent to which a set of measurements is free from random-error variance" (Guion, 1965, p. 30). The common rating errors just discussed are only one of several factors influencing the reliability and subsequently the validity of performance evaluations. First, methods of estimating reliability will be reviewed. Then, consideration will be given to additional factors affecting reliability.

Estimating Reliability

The reliability of ratings is traditionally assessed in one of two ways: (1) comparison of parallel ratings where two or more raters judge a single group of persons, or (2) comparison of two sets of evaluations obtained from the same rater on two different occasions. In both cases, the degree of agreement between the separate ratings of each ratee is taken as an indication of the precision or dependability of the measurements.

A high correlation between the parallel evaluations of two raters may indeed indicate that they have based their ratings upon the same criterion. A high reliability coefficient does not, however, insure high validity for the ratings as indicators of actual job performance. Although the parallel ratings may be based upon a common criterion, it is possible that this criterion is only one facet of total job performance. More serious yet, the ratings may have been based upon a totally irrelevant criterion. Therefore,

a high degree of conformity may indicate that the job performance sampled by the ratings is deficient. Conversely, it is possible for raters to make judgments based upon different yet valid criteria. In such a case the parallel ratings could possess a low reliability coefficient while obtaining an adequate sample of job performance. Buckner (1959) conducted a study whose findings illustrate this point. The subjective ratings of submarine sailors' performance were validated against aptitude tests and training success. The ratings were divided into four groups based upon their degree of interrater agreement. Test and training scores were more predictive of the less reliable ratings than of those with high agreement between raters. One explanation for this outcome is that the combination of discrepant ratings represented a greater proportion of total job performance variance. Such a conclusion argues for the usefulness of including, when possible, additional raters who are likely to observe different aspects of the ratees' performance.

Lawler (1967) suggests that evaluations of multiple traits be made from several vantage points and that the ratings be evaluated, through the use of the multitrait-multimethod (rater) matrix (Campbell & Fiske, 1959). He argues that this approach allows the researcher to gain a more sophisticated understanding of his criteria by assessing the discriminant and convergent validity of the ratings.

Kavanaugh, MacKinney, and Wolins (1971) recognize that the conventional approach to evaluating data using the

multitrait-multimethod matrix is inferential, implicit, and cumbersome. They therefore refined a technique capable of analyzing and summarizing multitrait-multimethod data in a more explicit, interpretable form. Specifically, they present an analysis of variance model which investigates four sources of variance:

1. Person variance--which indicates the agreement (convergent validity) of subjects over raters and traits.

2. Person by trait variance--which indicates the degree of discrimination of traits by raters (discriminant validity).

3. Person by rater variance--which indicates the amount of halo bias.

4. Error variance.

These authors argue that in addition to providing a broad basis for examining the reliability of ratings, the multitrait-multimethod scheme provides the only sufficient evidence for the content relevance of ratings.

The traditional alternative to estimating reliability with parallel ratings is re-rating by the same rater at a later time. It was noted that for parallel ratings a low degree of interrater agreement is not necessarily indicative of low validity for the ratings. This also holds true for reliabilities obtained through the re-rating method. Low correspondence to a previous rating may reflect actual changes in job performance. That substantial variations occur in workers' performance is a well-documented phenomenon

and this dynamic nature of job success presents a considerable source of error to the reliability of ratings taken over a period of time. A person's work performance may display a changing but consistent trend (as a result of learning, for example) or it may show irregular fluctuations (due to motivation, shift work, seasonal job requirements, personal problems, etc.). Ghiselli and Haire (1960) studied the performance of taxi drivers for over an 18-week period. Job success, in terms of volume of fares, was measured weekly. The mean of the intercorrelations for the 18 weeks was +.57. However, the correlation between first and last week fares was only +.19 and considerable changes had occurred in the rank order of the subjects over that period. When performance criteria were correlated with three preemployment tests, it was found that one test decreased in validity, a second test fluctuated, and the third test increased over the 18 weeks of the study. Ghiselli and Haire also reported, in the same study, a ten-year follow-up of the performance of a group of investment salesmen. The average increase in job success over that time period was as high as 650%, with no evidence of leveling off. In a similar study, Fleishman and Krutchner (1960) showed that different behaviors are required in the early and later phases of learning Morse Code. Bass (1962) has also shown that the relationships between the merit ratings of salesmen change over time.

In instances where re-ratings show low correspondence to previous ratings, it is difficult to determine whether

the ratings have been deficient and undependable, or whether they reflect the dynamic characteristics of the criterion over a period of time. It therefore seems that the most desirable procedure would be to obtain parallel and re-ratings. Determination of the discriminant and convergent validity, across raters and across time, would then provide a broader base for evaluating the utility of the ratings. Such a design, however, requires multiple raters making evaluations over an extended period of time. These conditions are, regretfully, not often possible in practice.

Blanz and Ghiselli (1972) provide an additional approach to the question of reliability of merit ratings. Coefficients obtained from parallel or re-ratings comparisons express reliability for the entire rating process including: the scales, ratees, and raters. Researchers are therefore unable to differentiate the relative contribution of any one of these three components to the overall reliability of the ratings. Ratings amenable to study according to the multi-trait-multimethod matrix provide an additional breakout of information concerning individual scales and raters stratified according to predetermined groups. Blanz (1965), however, has developed a rating procedure using scaling techniques which, he suggests, makes it possible to examine the individual reliability with which each ratee is evaluated, the reliability of each scale, and the reliability of the ratings made by each rater. Furthermore, this procedure does not require multiple raters or additional ratings on a

second occasion.

Blanz's technique, the mixed standard scale, is a variation of the standard rating scale in which the rater is presented with descriptions of varying levels of performance and required to select the one best describing the ratee. In the mixed standard scale the rater is presented with three descriptions representing high, average, or low performance for a particular trait. The rater is asked to evaluate each ratee with respect to each statement by indicating whether the ratee is better than the description, the description fits the ratee, or the ratee is worse than the description. The three descriptions for a particular characteristic can then be combined by the investigator and evaluated as a group. Descriptions for several traits are presented in a random order. Thus, the rater does not deal with all three descriptions related to one characteristic simultaneously. It is therefore unlikely that the rater will perceive the performance level of any particular description relative to the other two statements comprising that set.

The mixed scale procedure is, first, an effort to conceal from the rater the order-of-merit of the evaluations in an attempt to reduce response set errors; and secondly, a means for evaluating the consistency of rater's responses in terms of Guttman-like scaling properties. Below are the three descriptions comprising the "Diligence" dimension used in the Blanz and Ghiselli study.

High performance -- A real workhorse. He works much

harder than his job really requires.

Average performance -- He is sufficiently industrious and earnest in his work. You cannot accuse him of being lazy; nevertheless, you wouldn't say he is exceptionally diligent.

Low performance -- He has a touch of laziness. He does just what he is required to do, but no more.

By definition, the rater who rates correctly cannot indicate that more than one description in a set fits the ratee. Any statement which describes behavior superior to this accurate description must therefore be marked "ratee is worse than the description." Likewise, any statement describing behavior inferior to the accurate description must be evaluated "ratee is better than the description."

Using the foregoing scheme, Blanz and Ghiselli scored each set of three descriptions according to a 7-point continuum. Below is their scoring system for all possible combinations of error-free responses.

Descriptive Statements			Rating
High	Average	Low	
+	+	+	7
0	+	+	6
-	+	+	5
-	0	+	4
-	-	+	3
-	-	0	2
-	-	-	1

+ = ratee is better than the description
0 = description fits the ratee
- = ratee is worse than the description

All combinations of responses except the above seven are

considered illogical and in error. (The manner in which Blanz and Ghiselli score error free combinations will be presented and discussed later).

It is then possible, using the mixed standard scale, to identify the number of rating errors for a particular ratee, trait, or rater. Determination of the consistency of the ratings in the Blanz and Ghiselli study was based upon a form of scalogram analysis (Edwards, 1957; Forgerson, 1958).

In his doctoral dissertation, Blanz divided ratees into two groups based upon the magnitude of rating errors identified by the mixed standard scale. One group consisted of workers whose ratings contained fewer errors than the grand mean, while the other group was comprised of workers who had been rated with less consistency than the overall average. Correlation coefficients between test scores and ratings were then computed separately for the two groups. It was discovered that the correlations were significantly higher for the group whose ratings contained fewer errors. Blanz proposed that if the tests used in this study were valid predictors of job performance, it could be inferred that those ratings containing fewer errors were more dependable and relevant. Hence, the degree of error identified in the mixed standard scale ratings can be considered a measure of reliability and the validity of ratings can be increased through the removal of error ridden ratings. Based upon their research, Blanz and Ghiselli conclude that the

capability of the mixed standard scale to identify the extent of rating error for individual raters, ratee, and traits may in some cases make it superior to the more traditional methods of estimating reliability.

In summary, it has been shown that the reliability of performance measures can be assessed in a number of ways. Furthermore, these different methods can provide very different estimates of reliability in some cases. For this reason some researchers have concluded that a limit on the validity of a measure may not necessarily be set by its reliability value. Methods of estimating reliability should be based upon the goal of obtaining, within the practical limits of the situation, the broadest possible body of information for evaluating the utility of ratings. Toward this end, the multitrait-multimethod matrix and mixed standard scale techniques, as well as the more traditional parallel rating and rating-rerating comparisons, can contribute to our comprehension of the dynamic nature of job performance.

Additional Factors Influencing Reliability

Job related factors. In light of the dynamic nature of job performance, it would seem crucial to the reliability of evaluations for the rater to have the opportunity to observe the ratees at work and that observations be made over an extended period of time. Toward this end, Flanagan (1949) developed his critical-incidents technique of recording performance. This implies that the ratee has been in his job for a sufficient time period and that the rater knows

the ratee's job well enough to identify the critical tasks (Ferguson, 1949; Jurgensen, 1950b; Ward, 1961).

Just how much job experience is required for the ratee and how much opportunity for observation is needed by the rater are unresolved issues. Kliegen and Mosel (1953) found that excellent observation conditions do not always improve the reliability of ratings. For their study they formed two groups based upon whether the raters had little or considerable opportunity to observe the performance of their respective ratees. Contrary to expectations, no significant differences were found between the reliabilities of the ratings for the two groups. Regarding job experience, the complexity of tasks performed in the job would seem a prime determiner of the minimum period required before a job incumbent could be evaluated reliably. It appears pointless to evaluate an individual's performance before learning has reached at least an initial plateau or before the individual has encountered a sufficient variety of the situations typical to the job. However, the Ghiselli and Hare (1960) study illustrated that learning, and hence improvement in job performance, may never level off in some occupations.

The impact of employee seniority and skill level upon merit rating was investigated by Jay and Copes (1957). Four occupational groups were formed, based upon job skill levels: unskilled, semi-skilled, skilled, and professional. Validation studies using standardized methods had previously been conducted on all four groups. When job experience was

correlated with performance ratings in each group, it was found that the higher the skill level, the stronger the relationship between experience and performance. For the unskilled group, merit rating was totally independent of job experience ($r = -.05$, $N = 258$). In the skilled group the interdependence first reached a significant level ($r = .27$, $N = 642$). Thus, in this study skill and level of seniority were factors which influenced the results of the merit ratings.

It is, of course, possible that prolonged experience does in fact result in greater proficiency, especially in the more complex jobs and where novel situations continue to occur. It is a common observation, however, that subjective ratings tend to correlate with experience, even where more objective criteria fail to do so (Jay and Copes, 1957). Hausman and Stupp (1955), among others, report studies in which merit ratings correlated significantly with both job experience and age of the ratees. Efforts should therefore be made to statistically control for experience and skill level and to keep their effects separate from other results.

Rater characteristics. The quality of ratings as a function of the characteristics of the raters has also been the focus of a considerable body of research. Since individuals have been found to vary widely in their ability to judge others, it would be of great value to identify any specific characteristic of the evaluator having a predictable impact upon his judgment. Based upon a comprehensive

literature review, however, Guilford (1954) concluded that the reliability and validity of the ratings given by any one rater are likely to vary from situation to situation depending upon the particular rating traits and group of ratees. Guilford summarized the following characteristics as typical of good raters: has good insight into his own strengths and weaknesses; is secure and independent; has sympathy for others; is well adjusted; is an experienced rater and is aware of the common evaluation errors.

Intelligence of the raters has been found to be a factor affecting the reliability of raters (Stockford and Bissel, 1949). Intelligent raters not only rated more reliably than their less intelligent counterparts, but were also less biased by their length of acquaintance with the ratees.

Kirchner and Reisberg (1962) divided raters into two groups based upon the ratings given them by their superiors. The groups evaluated as better than average showed greater dispersion in their ratings of their own subordinates than did the group of less effective supervisors. The less effective group tended to avoid giving low ratings to their poorer subordinates. Furthermore, the raters seemed to reflect themselves in their ratings; above average supervisors giving better ratings to individuals high on initiative and independence while less effective supervisors place more importance on compliance and group cooperation.

The results of a study by Johnson and Vidulich (1956)

were somewhat in agreement with the idea that the rater places emphasis upon his own prominent personality traits. However, these authors point out that the rater's comparison of others to himself on a particular trait may have two opposite outcomes: because he perceives himself as possessing the trait he may rate others high as well, or he may feel that no one is his equal on the trait and rate others low by contrast.

The influence of evaluation training, and hence the rater's knowledge, has been studied to a limited degree. Such training may encompass several topics and serve numerous purposes. Certainly training should include instruction in rating techniques and the common sources of error. It is also important to provide the different raters with as uniform and objective a foundation as possible in using the rating instrument in order to make the ratings more nearly comparable.

Another important goal of training is to ensure the rater's motivation and favorable attitude toward the rating process. It may serve as a useful method of motivation to thoroughly explain the reasons why the ratings are being conducted, how the rater is likely to benefit, and to emphasize how decisively the quality of the ratings depends upon the rater. The raters' attitudes may be especially important when the number of persons to be rated is large or the rating process is complicated. In either case, fatigue is likely to make the ratings less reliable. Bayroff, Haggerty and

Rundquist (1954), found that the validity of ratings which required approximately one and one-half hours to complete declined somewhat during the last 45 minutes of the rating task. Bendig (1957) discovered that when the rating questions are complex or the rater knows there are a large number of traits to be rated, he is less likely to rate reliably.

Participation in the development of the rating items can precipitate both education for the raters in the use of the ratings and favorable attitudes toward the evaluation process. Use of the retranslation technique to develop performance criteria is one such means of getting the rater involved and more aware of the intricacies of the rating process. Training, in fact, may be a critical factor in the success of behaviorally based scales developed through the retranslation method. Hakel (1971), found that the use of such scales by untrained raters resulted in relatively low success.

That training raises the reliability of ratings was demonstrated by Stockford and Bissel (1949). They had two groups of raters--trained and untrained. The rating-re-rating reliability coefficient was higher for the trained than for the untrained group ($r = .85$ and $r = .76$ respectively). Furthermore, training had a greater effect upon the reliability of the evaluations of the more intelligent raters. The correlation between intelligence and rater reliability was .52 for the trained group and .20 for the

untrained group. Apparently the more intelligent raters are better able to grasp and apply instruction which is given them.

In summary, the studies reported in the literature suggest that significant inter-individual differences occur in terms of the quality and usefulness of ratings. But there does not appear to be evidence for a general judicial ability. Instead, the quality of ratings is dependent upon numerous characteristics of the rater and upon the circumstances under which the rating occurs. In light of these conclusions, it seems especially useful to obtain performance observations from raters at different vantage points and to study these data according to the multitrait-multimethod matrix.

Rating form characteristics. Reliability as a function of the length of the rating form was investigated by Taylor, Schneider, and Clay (1954). Surprisingly, they were able to attain a two-thirds reduction in the length of a forced-choice rating form without a significant reduction in the reliability of the results. It would seem that the danger in such a reduction lies in the potential for eliminating some of the traits critical for measuring job success.

Research has not disclosed an optimal number of rating categories ensuring high reliability. Bendig (1954) experimented with rating scales where the number of categories varied from two to nine. No clear-cut superiority appeared for any particular number of categories. Matell and Jacoby

(1972) varied the number of Likert scale steps from two to 19 for purposes other than the examination of reliability. They found that mean response time increased while the usage of an "uncertain" category decreased as the number of rating steps increased. Guilford (1954) reviewed a number of studies and concluded that the number of categories should exceed seven. Guilford felt there might even be situations where as many as 25 rating classes could be advantageous. Guion (1965) disagrees with these conclusions, stating it is unlikely that most raters can successfully discriminate more than seven categories. However, Guion admits that empirical evidence is lacking.

The relative location of scales on a rating form has been discovered to result in a proximity effect. Stockford and Bissel (1949) were the first to find that two traits adjacent to one another on the rating form tend to correlate more highly than any two traits placed farther apart. In their study they found that the average correlation between two traits diminished from .66 to .46 as the number of traits separating them increased. An alternative to decreasing the rating form space between two traits might be to increase the elapsed time between rating any two traits. This can be achieved by requiring the rater to rate every individual on one trait before proceeding to the next trait. This technique was previously discussed as a method of reducing halo effects.

Determination of the ideal number of traits which can

be rated reliably appears to be an unattainable goal. Certainly, the concept of a single criterion of job success is unrealistic. There seems little doubt that job success consists of a number of rather independent factors and therefore, one needs to think in terms of partial criteria. Strong cases for multiple criteria have been made by Otis (1940); Toops (1944); Ghiselli (1956); Guion (1961); Weitz (1961); Dunnette (1963); Biesheuvel (1965); Guion (1965); Wallace (1965); Ronan and Prien (1966); and Roach and Wherry (1970) among others. The evidence from factor-analysis studies (Ewart, Seashore & Tiffin, 1941; Grant, 1955; Rush, 1953) indicates approximately three to five factors should be obtained from the various scales used in an evaluation.

The related question of the kinds of traits upon which evaluations are to be obtained is also complex. Probably one rating that should consistently be included is a global evaluation of effectiveness. Such a rating has merit in that it provides an overall index of job success. Furthermore, it may help to reduce the halo effect by allowing raters to get their positive feelings toward the raters documented and then get on with the task of evaluating independent components of performance. Campbell, Dunnette, Lawler and Weick (1970) present an impressive plea for allowing the individuals who will be performing the evaluations to determine the job behavior domain and to define these performance dimensions according to their own jargon. These authors oppose what they feel is "a common tendency

for psychologists to impose their own beliefs about job behavior and their own systems for recording it upon the persons whose task it is to observe that behavior." (pp. 118-119). They therefore advocate the determination of the behavior domains through the use of the critical incident-retranslation approach. As discussed previously, considerable success in obtaining independent performance factors has been achieved with this approach.

Statement of the Problem

Summarizing the review of literature the question might be asked, which of the various evaluation methods is superior? The most advantageous method depends, of course, on many factors, which vary with each particular situation. These factors include: characteristics peculiar to the raters, ratees, and job(s) in question; the opportunities and costs related to collecting data; the purposes for which the rating results are intended; and so on.

The ranking methods are relatively easy and hence inexpensive to construct and they eliminate the distribution problems of skewness and central tendency. However, these methods lend themselves better to global evaluation, and do not indicate the magnitude of inter-individual differences. Furthermore, the rating task becomes progressively laborious and less accurate as the number of individuals to be ranked increases.

Perhaps the greatest advantage of the forced-choice and

checklist methods is that the rater is not aware of the final result of the evaluation and it is more difficult to give higher ratings. In a study by Taylor and Wherry (1951) both leniency and central tendency errors were less for ratings using the forced-choice method than for those employing the graphic scale method. On the negative side, the forced-choice and checklist methods are also more appropriate to a global effectiveness measure and in addition the amount of rating and administrative work is higher than for the ranking or rating methods. Furthermore, raters tend to object to secret scoring procedures and sometimes are able to obtain the scoring key.

Rating scale methods are the most popular. They enable measurement of inter-individual differences on a variety of traits. Furthermore, in their simplest form they are quick and inexpensive to construct and analyze. However, the quality of rating scales seems to be inversely related to the ease of their construction. The simplest numerical and graphic scales are the most susceptible of all the evaluation methods to halo, this effect having been found to be smaller when ratings are not obviously made on a scale (Ewart, Seashore & Tiffin, 1941; Sisson, 1948). When painstaking efforts have been applied to the development of behaviorally based rating scales, the results have been encouraging. Barrett, Taylor, Parker, and Martens (1958) and Peters and McCormick (1966) have found formats incorporating behavioral descriptions to be of superior reliability compared with

numerically anchored scales. Similarly, Mass (1965) has found rater agreement on scaled-expectation ratings to be significantly superior to that of adjective rating scales. In a direct test between behaviorally based scaled expectations and less specifically defined general scales, Campbell, Dunnette, Arvey and Hellervik (1973) found that the scales defined by specific behavioral anchors yielded less method variance, less halo error, and less leniency error. Given, then, that highly specific behavioral anchors improve rating scales, it becomes a question of whether the higher quality of ratings offset the increased time and expense required to develop these scales.

Blanz and Ghiselli (1972) also utilize behavioral descriptions for their mixed standard scale technique which was described previously. Blanz (1965) developed the mixed standard scale with the intention of reducing halo and leniency errors by concealing the final result of the evaluation from the rater. Concealment has been found effective in reducing these errors using the forced-choice format. Blanz and Ghiselli concluded from the result of their study that the mixed scale ratings did not possess a large leniency error and factor analysis yielded four clearly distinguishable factors. It is especially noteworthy that the behavioral descriptions used in their mixed scales were not developed using the retranslation technique and were not nearly as specific as those employed in the Campbell, Dunnette, Arvey and Hellervick study. Yet the results obtained in these two

studies seem comparable.

It would therefore appear that the researcher has at least two potentially effective methods of reducing the common evaluation errors. Critical incidents may be collected and retranslated to develop scales with highly specific anchors. Or, less behaviorally specific performance descriptions may be randomly presented in a way which conceals from the rater the final order of merit that will be assigned to the ratings. A test of the relative effectiveness of these two rating methods in reducing halo, leniency, and central tendency errors has not been previously conducted. Such a test would not reveal the relative contributions of specificity of the anchors and concealment of the scale continuum, however, since both factors would have varied. The more appropriate design, presented in the 4-fold table of Figure 1, requires 4 rating methods. In this manner specificity of behavior can be controlled while scale continuum is varied and conversely, scale continuum controlled while specificity of behavior is varied.

Comparisons between various pairs of the cells presented in Figure 1 have been made. As reported previously, the Campbell, Dunnette, Arvey, and Hellervik (1973) study compared behaviorally specific and general scales both presented on obvious continuums (cells I and III). The results of their study were strongly in favor of the behaviorally specific scales.

Arvey and Holye (1974) made a direct test of highly

Specificity of Behavioral Anchors	Scale Continuum: Obvious	Scale Continuum: Non-Obvious
Specific	I	II
General	III	IV

Figure 1. Experimental design to compare the relative effects of scale continuity and specificity of behavior.

specific behavioral anchors presented in standard and mixed scale formats (cell I vs. cell II). Two sets of comparable scales were developed according to the critical incident-retranslation technique. On one scale the behavior incidents were arranged on a typical 1 to 7 point scale. The second type of scale was constructed for each dimension by arranging the incidents in random order on one page. Raters used both formats to rate their subordinates. They responded to the standard set of ratings by assigning a value from 1 to 7 for each dimension. For the mixed scale, raters were instructed to read each incident and decide whether the individual he rated was the same as, better than, or not as good as the individual described in the incidents. The results of these two ratings, when analyzed according to a multitrait-multimethod matrix were almost identical. Both methods demonstrated acceptable convergent and discriminant validity. The results were only slightly in favor of the standard, more obvious order-of-merit scale. Therefore, with specificity of anchors controlled, disguising the continuity of the rating scale did not improve the ratings, at least in terms of convergent and discriminant validity. The means and standard deviations of the ratings, unfortunately, were not factor analyzed.

The Blanz and Ghiselli (1972) study examined only the characteristics of scales utilizing behaviorally general statements presented on a non-obvious continuum (cell IV). Blanz and Ghiselli did not compare their mixed scale with a

scale representative of any other cell in Figure 1.

The present study employs, with one exception, the research design presented in Figure 1. Using this design, the relative effects of both scale continuity and specificity of behavioral anchors can be assessed. Rating scales were constructed to represent cells I, III, and IV of Figure 1: an obvious continuum scale anchored by behaviorally specific descriptions of performance (cell I); the same obvious continuum scale anchored by more behaviorally general performance descriptions (cell III); and behaviorally general descriptions presented in the non-continuous mixed standard scale format (cell IV).

A rating scale representing the behaviorally specific and non-continuous scale characteristics of cell II was not attempted in the present research. Responses to behaviorally specific statements arranged in random order may be scored by Guttman scaling techniques as in the Arvey and Hoyle (1974) study. The same behaviorally specific statements, however, do not lend themselves logically to the response and scoring system of the mixed standard scale method. To illustrate, the behaviorally general statement which describes a moderate level of performance can contain both a positive and negative component. The behaviorally specific statement which anchors the middle range of a scale does not contain this double frame of reference, however. The rater is therefore likely to say that the moderate level of performance described by a behaviorally specific anchor fits an outstanding employee.

Such a response would be considered consistent with Guttman scaling but would be an error response according to the mixed standard scale scoring method. For this reason, behaviorally specific anchors were not utilized in mixed standard scale format. Only scales representing the three remaining cells (I, III, and IV) of Figure 1 are administered and compared.

CHAPTER III

METHOD

Experimental Design

The purpose of this research was to contrast the psychometric properties of three types of merit rating scales: Behaviorally specific scales, behaviorally general scales, and behaviorally general-mixed standard scales. These three types of scales differ in the following respects. The behaviorally specific scales consist of explicit behavioral examples developed according to the critical incident-retranslation technique and presented as anchors along a typical 7 point continuous scale. The behaviorally general scales also employ the typical 7 point continuous scale format. However, the behavioral anchors used in these scales are general descriptions rather than explicit examples of behavior. The behaviorally general-mixed standard scales utilize the same behavioral descriptions contained in the behaviorally general scales. These descriptions are then mixed in random order and presented to raters in a noncontinuous scale format.

This rating scale research was incorporated in a validation project conducted to validate a pre-employment test battery for the selection of retail department store managers. The rating scales were developed to assess the job performance of incumbent store managers. The department store chain is divided for supervision into 5 regions. Each region is further divided into 5 to 6 districts. The district

(first line) managers and regional (second line) managers assisted in development of the scales and provided the performance evaluations.

The experimental design of the study is presented in Table 1. In each condition the merit ratings obtained using two of the three types of rating scales were compared: Behaviorally general with mixed standard, behaviorally general with behaviorally specific, and mixed standard with behaviorally specific. District managers were assigned to rating conditions so that an approximately equal number of store managers would be rated with each pair of the three combinations of methods (Table 1). Furthermore, at least one district supervisor from every region was assigned to each rating condition. Regional supervisors and their respective district supervisors rated the same store managers using the same rating method. This means that on the first rating occasion, regional supervisors used the mixed standard scale to rate some store managers and the behaviorally general scale to evaluate the remainder. On the second rating occasion this procedure was repeated using instead the behaviorally general and behaviorally specific scales, as appropriate for each store manager.

Scale Development

Behaviorally Specific Scales

Formulation of incidents. An initial series of interviews were conducted with a cross-section of the regional

TABLE 1

Experimental Rating Conditions

	Number of First Line Raters[a]	Number of Ratees	Scale Used On First Rating Occasion	Scale Used On Second Rating Occasion
Condition 1	8	60	Behaviorally General-Mixed Standard Scale	Behaviorally General Scale
Condition 2	7	53	Behaviorally General Scale	Behaviorally Specific Scale
Condition 3	8	64	Behaviorally General-Mixed Standard Scale	Behaviorally Specific Scale

[a]The five second line raters evaluated all their respective ratees in each of the three conditions.

and district supervisors who later evaluated store managers. These supervisors were asked to describe critical incidents of store manager behavior, with no prior discussion of the underlying performance factors. From these descriptions a list of performance categories was developed and tentative definitions for each category were written. These definitions were then presented to a different group of supervisors in a second series of interviews. The latter group of supervisors was asked to comment on the clarity, relevance, and inclusiveness of the tentative factors and to provide specific behavioral incidents representing the various performance factors. Following these discussions performance dimensions were added, combined, and dropped from the original list with the result that 18 performance categories were finally retained. For each dimension retained there were a minimum of three behavioral incidents which appeared to represent moderate as well as extremely effective and ineffective levels of performance. This final list was presented to company personnel who edited the statements to insure that the language used was organizationally correct.

Retranslation procedure. The retranslation step of the Smith and Kendall procedure was carried out next. A list of performance categories, including their definition, was mailed to each supervisor participating in the rating process, along with the behavior incidents intended to represent those categories. In order to reduce the workload of the raters, each of the 18 performance categories was randomly assigned

to one of two lists and each rater received only one-half of the total number of categories. The participants were asked to make three judgments. First, each evaluator read a list of performance dimension definitions and rated the importance of each on a 5-point scale; 5 indicating behavior which is "absolutely vital to overall effectiveness" as a store manager and 1 indicating "makes no difference whatsoever" in managing a store. Next, the list of corresponding specific behavior incidents was presented and the supervisors were asked to sort each incident into the dimension that it most closely represents. Finally, the supervisors rated each incident on a 7-point scale indicating its degree of effective or ineffective behavior on the performance dimension in which it was placed.

Specific behavior incidents were retained and used in scales only which there was high (80% or better) agreement among the judges as to the incident's performance category and when the standard deviations of the effectiveness values assigned by the judges were comparatively small. Those performance categories which did not clearly retranslate or which did not have suitable anchors dispersed along the scale were excluded from the study.

The retranslation procedure resulted in the elimination of seven of the original scales. The eleven retranslated dimensions plus an overall effectiveness dimension which was added as a final comprehensive scale are presented in Table 2. Table 3 displays the mean, standard deviation and

TABLE 2

Performance Dimensions Used in All Rating Methods

1. PLANNING AND ORGANIZING: Ability to plan, organize and schedule current and future work so that it is performed in the proper sequence and her time is effectively utilized.

2. UNDERSTANDING AND CARRYING OUT INSTRUCTIONS: Ability to comprehend and carry out verbal or written instructions and directions correctly without requiring additional information or supervision.

3. WORK EFFORT: Putting in extra time and effort, doing more than is demanded by higher management.

4. SELECTING NEW EMPLOYEES: Identifying and hiring the person most likely to be a good employee.

5. EFFECTIVE USE OF EMPLOYEES: Making work assignments in a way which best takes advantage of the abilities of her employees.

6. REWARDING AND MOTIVATING EMPLOYEES: Giving praise when it is deserved, letting employees know their efforts are noticed and appreciated.

7. CUSTOMER RELATIONS: Greeting and serving customers, attending to their needs, and handling their complaints in a manner which keeps them satisfied and coming back.

8. STORE CLEANLINESS AND APPEARANCE: Keeping the store as clean and attractive as is possible, considering the conditions at that particular store.

9. CONTROL OF SHRINKAGE: Keeping shrinkage to a minimum, considering the special circumstances at that particular store which affect shrinkage figures.

10. ACCURACY OF WORK: Making few or no errors in completing sales reports, doing markdowns, etc.

11. SALES: Motivating customers to buy merchandise.

12. OVERALL EFFECTIVENESS: Level of effectiveness in performing the duties of her current position as store manager.

TABLE 3

Results of Retranslation Procedure Performed on Behavioral Anchors for Behaviorally Specific Scales

Performance Dimension	Anchor	Mean Level of Performance	SD	% Correct Retranslation
1. Planning and Organizing	high	5.30	.67	100
	middle	4.20	1.03	100
	low	2.30	.82	100
2. Understanding and Carrying out Directions	high	5.45	.68	100
	middle	4.18	.60	90
	low	2.27	.64	100
3. Work Effort	high	5.59	1.15	85
	middle	4.72	1.05	85
	low	2.45	.82	80
4. Selecting New Employees	high	5.64	1.02	90
	middle	4.27	1.00	90
	low	1.72	.64	100
5. Effective Use of Employees	high	5.36	.80	100
	middle	4.54	.93	90
	low	2.45	.93	100
6. Rewarding and Motivating Employees	high	4.80	1.03	100
	middle	3.80	.42	100
	low	2.10	.73	90
7. Customer Relations	high	5.70	.82	100
	middle	4.10	.31	90
	low	2.00	.81	100
8. Store Cleanliness and Appearance	high	5.36	.92	80
	middle	4.18	.98	100
	low	3.00	.77	90
9. Control of Shrinkage	high	5.90	1.10	100
	middle	4.30	.82	90
	low	2.40	1.34	80
10. Accuracy of Work	high	5.40	.81	100
	middle	3.50	.52	100
	low	2.50	1.26	100
11. Sales	high	5.90	1.10	100
	middle	4.00	.00	100
	low	2.20	.42	100
12. Overall Effectiveness	Not Retranslated			

percentage of correct retranslations for each behavioral anchor. The final scale for each of the dimension consisted of the scale name, scale definition, and a 7-point continuum defined by specific behavior incidents located adjacent to the appropriate scale values. Each specific behavioral incident was stated in the form "could be expected to..." in order to avoid the domain sampling problem associated with implying that the ratee must actually exhibit that specific behavior.

Behaviorally General Scales

Time restraints imposed by the client organization prohibited analysis of the retranslation data prior to construction and administration of the general and mixed scales. For this reason, general behavioral statements were developed to represent all of the performance dimensions defined and submitted to the retranslation technique. Those seven dimensions eliminated from the behaviorally specific scales due to a failure to retranslate were then eliminated from the analysis of the behaviorally general and behaviorally general-mixed standard scales.

Appropriate general statements for the behaviorally general scales were derived from the definitions for the behaviorally specific scales. The major elements of the scale definition was incorporated in each of three statements and worded so as to illustrate highly effective, effective but not outstanding, and ineffective performance

levels for that particular scale. These three statements were located at points 6, 4, and 2, respectively, on an obvious 7-point continuum. To form statements describing highly effective performance, positive modifiers were attached to the basic definition. In a like manner, negative modifiers were inserted to describe ineffective performance. The statements describing average effectiveness levels contain a combination of both positive and negative qualifiers. The procedure is illustrated with the example in Table 4. The proposed performance dimension is titled "Rewarding and Motivating Employees." The definition precedes the general statements describing effective, average, and ineffective levels of performance. For comparison purposes, behaviorally specific statements have been provided adjacent to the more general descriptions of performance. These retranslated specific statements anchor approximately the same scale values on the specific scales as those arbitrarily assigned to the general statements on the general scale (i.e., levels 6, 4, and 2 on the 7-point scale).

For comparative purposes, it should be noted that the performance dimension used in the rating are identical for the behaviorally specific and behaviorally general scales--those dimensions not surviving the retransaltion step for the specific anchors having also been eliminated from the behaviorally general scales. The difference between the behaviorally specific and general scales is specificity of the behavioral anchors. That is, specific anchors provide

TABLE 4

Comparison of Behaviorally General and Behaviorally Specific Scale Anchors

Performance Dimension: Rewarding and Motivating Employees

Definition: Giving praise when it is deserved, letting employees know their efforts are noticed and appreciated

	General Anchors	Specific Anchors
Highly Effective Performance	This manager looks for every opportunity to praise her employees and to show that their efforts are noticed and appreciated.	If an employee performed an assigned task even better than expected, she might say, "That's a very good job, even better than I could have done."
Average Performance	While this manager may not always tell her employees when they have done an exceptionally good job, she sometimes lets them know that she is pleased with their work.	When leaving for the night, she could sometimes be expected to say, "Girls, we've accomplished a lot today. I feel you've all done a good job."
Ineffective Performance	This manager takes her employees for granted. She seldom tells them when they have done a good job.	She could be expected to use a good suggestion made by an employee without giving her credit for it.

behavioral examples of the definition for the scale which they represent, while the general anchors are restatements of that same scale definition, worded so that they represent high, average, and low levels of that dimension. Both the behaviorally specific and general anchors were presented to the raters on an obvious order-of-merit continuum (a 7-point scale).

Behaviorally General-Mixed Standard Scales

The descriptive statements for these scales are identical to those anchoring the behaviorally general scales. However, the mixed standard scale statements for all performance dimensions were presented to the raters arranged in random order, thereby reducing the likelihood that the rater would perceive an order-of-merit continuity between any one statement and the two remaining statements for that dimension. The statements were then reordered to reflect a continuum from high to low effectiveness and scored on a 7-point basis comparable to the behaviorally specific and behaviorally general scales.

For the mixed standard scales, the ratee's performance level relative to each description is indicated by a plus (manager is better than the description), zero (description fits the manager perfectly), or minus (manager is not as good as the description). The evaluations were recorded on pages containing the names of all store managers to be evaluated by that particular rater.

The basis for assigning rating scores to error responses on the mixed standard scales was somewhat different from that proposed by Blanz and Ghiselli (1972). In Table 5 the scoring system used in the present study for illogical responses is presented along with the system employed by Blanz and Ghiselli.

Blanz and Ghiselli interpret an illogical response combination as a judgment error on the part of the rater. An alternative explanation for an illogical response is that the rater made a placement error, either in reading the descriptive statement or in recording his response. Examination of an inconsistent set of responses in light of these two possible sources of error requires selection of a scale value from a range of potentially "true" values. A scoring system was therefore selected which is likely to result in the least variance in relation to the non-error combinations most logically derived from the error response set.

The decision rules for assigning scale values to inconsistent responses were based upon the number of conflicting responses within any given 3-statement combination. When one response is in conflict with the remaining two, the combination is scored as though the conflicting response had been changed to agree with the other two. The first four response combinations in Table 5 illustrate this rule. In each combination the low statement has a conflicting response and each combination has been scored as it would had the low response not been in conflict.

The fifth, sixth and seventh response combinations

TABLE 5

Contrasting Values Assigned to Illogical Responses in the Blanz and Ghiselli (1972) Study and in the Present Research

Response Combination			Scale Value Assigned		
High	Average	Low	Present Study		Blanz and Ghiselli Study
+	+	0	7		7
+	+	-	7		7
0	+	0	6		6
0	+	-	6		6
+	-	+	5	X	3
+	0	+	5	X	4
0	0	+	5		5
+	0	0	4	X	3
+	0	-	4		4
0	0	0	4		4
0	0	-	4		4
0	-	+	4	X	5
-	+	0	4	X	5
-	+	-	3	X	5
-	0	0	3		3
-	0	-	3		3
+	-	0	2		2
0	-	0	2		2
+	-	-	1		1
0	-	-	1		1

+ = ratee is better than this statement
0 = statement fits the ratee
- = ratee is worse than this statement

X Indicates disagreement between the two scoring methods

present in Table 5 illustrate the case where two responses are in conflict with one another, yet the third response agrees with both. The rule for this condition produces an average between the two non-error values obtained when each of the conflicting responses is corrected to agree with the other two. For example, in the fifth response combination in Table 5, the high and average responses are in conflict. If the high response were changed to a minus, the resulting combination would be logical and assigned a scale value of 3. On the other hand, changing the middle response to a plus would also result in a logical combination which would be assigned a scale value of 7. The average between these two extremes of 3 and 7 is a scale value of 5 which has been assigned to this illogical combination. In those cases where the average of the two extremes is not an integer, the value has been rounded toward the center of the 7-point scale.

The third decision rule involves those instances where all three responses of a set are in conflict with one another. Combination numbers 8, 9, 10, and 11 in Table 5 comprise the group of sets containing three responses which are all in conflict. For these cases, the decision rule is to assign the combination a scale value of 4--the center of the 7-point scale.

Subjects

Raters: First and second line supervisors in a department store chain performed the ratings for this study. The first-line raters consisted of 23 district supervisors, each

responsible for 6 to 12 department stores. The second line raters were the 5 regional supervisors, each responsible for 5 to 6 districts. With one exception, all raters were male.

Ratees: 177 managers of retail department stores located in small towns throughout the southeast and south-central United States were evaluated on their job performance. All ratees were female with a minimum of six months tenure as store managers.

Scale Administration

As mentioned previously, the ratings for this research were part of a test validation study. The store managers were administered a test battery as the initial phase of this project. This battery was administered, but not scored, by the district managers. Each store manager was rated on two separate occasions by her first- and second-line supervisors using different rating methods for each occasion. The experimental design of the study involves comparing each of three rating methods with the other two. Thus, for the two rating occasions, raters used one of the following combinations of scales: behaviorally specific and general, general and mixed, and mixed and behaviorally specific.

There were, however, two limitations upon the manner in which the experimental conditions could be varied. First, time constraints prevented the use of the behaviorally specific scales on the first rating occasion, since the retranslation procedure extended development of the specific scales past the deadline for conducting the initial ratings. Secondly,

it seemed preferable that those raters using the mixed scale not see the underlying performance dimensions prior to rating with the randomly mixed descriptions. Such would be the case if raters used either the specific or general scales before rating with the mixed scales. Therefore, the mixed rating scales were used only on the first rating occasion.

Because it was not possible to provide the supervisors with training concerning performance evaluation, the rating instructions were prefaced with a brief description of the purpose of the study and suggestions related to avoiding the common evaluation errors (Appendix A is an example of the instructions for rating with the mixed scale).

The instructions for the behaviorally specific and behaviorally general scales directed the raters to read one scale and then to evaluate all their store managers on that dimension before proceeding to the next scale. Instructions for the mixed scale asked the raters to follow the same procedure in reading one performance description and then rating all managers before going on to the next description.

The behaviorally specific and behaviorally general scales for the second set of evaluations were not mailed to the raters until all the behaviorally general and mixed standard scale evaluations from the initial rating occasion had been completed and returned. The elapsed time between the mailing of the first and second evaluations was 9 weeks.

CHAPTER IV

RESULTS AND CONCLUSIONS

Scalability of the Mixed Standard Scales

The first analysis involved determining the extent to which the mixed scales exhibit adequate scaling properties. The reproducibility of each scale was computed using the formula proposed by Guttman (1944):

$$\text{coefficient of reproducibility} = 1 - \frac{\text{number of errors}}{\text{total number of responses}}$$

Guilford (1954) suggested that reproducibilities approaching .80 represent quasi scales while Guttman considered reproducibilities over .85 to indicate that the data are sufficiently scalable.

Error responses were computed by two separate methods. First, the scales were scored and reproducibility coefficients were calculated through a variant of Guttman's "Cornell technique" proposed by Goodenough (Edwards, 1957, pp. 184-197). With this technique errors are identified based upon the actual responses of a specific sample. As a second procedure, the reproducibility coefficients were calculated using response errors based upon a "logical" response scheme developed in the present study. Table 6 contrasts the scoring and error count procedures when using the Goodenough and "logical" response schemes. For this particular group of ratings the scale scores for the two methods were identical. However, in a few cases, the error counts were quite different between the two.

The coefficients of reproducibility obtained for each

TABLE 6

Scoring and Error Count Procedures for the Goodenough and "Logical" Scaling Methods

Response Combination			Scale Score	Number of Response Errors	
High	Average	Low		Goodenough	"Logical"
+	+	+	7	0	0
+	+	0	7	1	1
+	+	-	7	1	1
0	+	+	6	0	0
0	+	0	6	1	1
0	+	-	6	1	1
-	+	+	5	2	0
+	0	+	5	1	1
+	-	+	5	2	1
0	0	+	5	0	1
-	0	+	4	2	0
0	-	+	4	2	1
-	+	0	4	2	1
+	0	0	4	1	2
+	0	-	4	2	2
0	0	0	4	0	2
0	0	-	4	1	2
-	-	+	3	2	0
-	+	-	3	2	1
-	0	0	3	0	1
-	0	-	3	1	1
-	-	0	2	0	0
+	-	0	2	2	1
0	-	0	2	1	1
-	-	-	1	0	0
+	-	-	1	1	1
0	-	-	1	1	1

+ = ratee is better than this statement
0 = statement fits the ratee
- = ratee is worse than this statement

scale using the Goodenough and "logical" error methods are presented in Table 7. For the Goodenough method only one scale, "Work Effort", failed to meet Guttman's stringent scalability criterion of .85, while all twelve scales exceeded Guilford's more lenient criterion of .80. Conversely, only four of the twelve scales exceeded a reproducibility of .80 when using the "logical" method of counting response errors, and none of the scales reached the .85 level. Blanz (1965) obtained a similar discrepancy between the reproducibility coefficients computed using the two methods of identifying errors.

Comparison of Mean Ratings

Leniency effects were evaluated by comparing the mean ratings on the 12 dimensions when the same raters used two different rating methods. The mean for each dimension was computed separately for first and second line supervisors. These results are presented in Tables 8 and 9 respectively. As approximate tests of the significance of differences between methods, student's t-tests using difference scores (Hays, 1973, pp. 335) were performed on the means obtained using each pair of rating methods. Significant differences were obtained in only one rating condition for both first and second line supervisors. For first line supervisors the means of the behaviorally general scales were significantly lower ($p < .05$) than those of the behaviorally specific scales; while for second line supervisors the means of the behaviorally general-mixed standard scales were significantly lower ($p < .01$) than those

TABLE 7

Reproducibility Coefficients of the Behaviorally General-Mixed Standard Scales

	Performance Dimension	Coefficient of Reproducibility: Goodenough	Coefficient of Reproducibility: Logical
1.	Planning & Organizing	.93	.73
2.	Understanding & Carrying Out Instructions	.93	.74
3.	Work Effort	.83	.80
4.	Selecting New Employees	.93	.84
5.	Effective Use of Employees	.89	.81
6.	Rewarding & Motivating Employees	.92	.66
7.	Customer Relations	.96	.83
8.	Store Cleanliness & Appearance	.93	.74
9.	Control of Shrinkage	.92	.81
10.	Accuracy of Work	.94	.61
11.	Sales	.89	.76
12.	Overall Effectiveness	.91	.79

TABLE 8

Mean Ratings on 12 Performance Dimensions Rated by 1st Line Supervisors

Experimental Condition	Performance Dimension	Mean Rating											
		1	2	3	4	5	6	7	8	9	10	11	12
Condition 1 t = .21	Behaviorally General-Mixed Standard Scale	4.38	4.67	4.74	4.82	5.30	4.62	5.32	3.73	4.65	4.40	4.62	4.80
	Behaviorally General Scale	4.60	4.57	4.78	4.77	4.62	4.82	4.85	4.58	4.33	4.57	4.70	4.57
Condition 2 t = 2.81*	Behaviorally General Scale	4.15	4.28	4.75	4.25	4.60	4.43	5.04	4.23	4.60	4.49	4.58	4.60
	Behaviorally Specific Scale	4.36	4.49	5.11	4.92	4.83	4.40	5.17	4.40	4.89	4.92	4.23	4.70
Condition 3 t = .85	Behaviorally General-Mixed Standard Scale	4.58	4.98	5.36	4.94	4.89	5.08	5.05	4.56	5.19	4.41	4.75	4.70
	Behaviorally Specific Scale	4.47	4.83	4.84	4.95	4.69	4.34	5.22	4.81	4.67	5.19	4.45	4.83

* $p < .05$

TABLE 9

Mean Ratings on 12 Performance Dimensions Rated by 2nd Line Supervisors

Experimental Condition	Performance Dimension	Mean Rating 1	2	3	4	5	6	7	8	9	10	11	12
Condition 1 t = 5.53*	Behaviorally General-Mixed Standard Scale	3.97	4.55	4.38	5.20	4.85	4.35	5.47	3.63	4.02	3.95	3.87	4.42
	Behaviorally General Scale	4.97	4.72	5.33	4.88	4.88	4.98	5.65	4.60	4.35	4.97	4.67	5.00
Condition 2 t = .42	Behaviorally General Scale	5.02	5.62	5.70	5.45	5.28	5.23	5.96	5.04	5.19	5.57	5.38	5.36
	Behaviorally Specific Scale	5.45	5.32	5.84	5.15	5.49	5.30	5.60	5.19	5.28	5.52	5.47	5.57
Condition 3 t = .97	Behaviorally General-Mixed Standard Scale	4.08	4.55	4.53	5.30	5.44	4.52	5.19	3.84	4.70	4.08	3.80	4.63
	Behaviorally Specific Scale	4.73	4.90	4.75	4.67	4.81	4.28	4.78	4.61	4.78	5.00	4.48	4.69

* $p < .01$

of the behaviorally general scales. It was concluded that there were no consistent differences among the three methods as far as leniency errors were concerned.

Comparison of the Variability of the Ratings

Variability of the three rating methods was compared by examining the magnitude of the standard deviations of the ratings obtained for each of the three rating conditions. As in the comparison of leniency effects, the data were analyzed separately for first and second line supervisors. Table 10 presents the standard deviations of the first line raters. Based upon approximate t-tests using difference scores, significant differences in the standard deviations of ratings by first line supervisors on the 12 dimensions were obtained for all three conditions. The mixed-standard rating method produced slightly but consistently larger standard deviations than did the other two methods when the ratings were made by first line supervisors. Standard deviations of the behaviorally general scales were slightly larger than for scales with specific scale anchors. On the other hand, as can be seen from Table 11, there were no significant differences in the standard deviations of the three rating methods on ratings made by second line supervisors.

There are several possible explanations for these findings. Since the rating method obtaining the highest standard deviations was always used first, there may be an order effect for first line supervisors. Some evidence on this issue can

TABLE 10

Standard Deviations of 12 Performance Dimensions Rated by 1st Line Supervisors

Experimental Condition	Performance Dimension	Standard Deviation											
		1	2	3	4	5	6	7	8	9	10	11	12
Condition 1 t = 3.16*	Behaviorally General-Mixed Standard Scale	1.49	1.13	1.46	1.38	1.01	.94	1.23	1.68	1.74	1.18	1.40	1.39
	Behaviorally General Scale	1.09	1.18	1.01	1.09	1.24	1.07	.95	1.25	1.46	1.08	1.05	1.03
Condition 2 t = 2.42*	Behaviorally General Scale	1.49	1.29	1.22	.83	1.48	1.03	1.29	1.41	1.36	1.44	1.38	1.23
	Behaviorally Specific Scale	1.33	1.20	1.30	1.19	1.07	.91	1.01	1.12	1.20	1.09	1.10	1.17
Condition 3 t = 7.11**	Behaviorally General-Mixed Standard Scale	1.56	1.37	1.25	1.63	1.40	1.06	1.45	1.59	1.42	1.29	1.40	1.52
	Behaviorally Specific Scale	1.23	1.16	1.05	1.25	1.11	1.14	1.11	1.23	1.15	1.01	1.23	1.18

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

TABLE 11

Standard Deviations of 12 Performance Dimensions Rated by 2nd Line Supervisors

Experimental Condition	Performance Dimension	Standard Deviation											
		1	2	3	4	5	6	7	8	9	10	11	12
Condition 1 t = .06	Behaviorally General-Mixed Standard Scale	1.36	1.24	1.33	1.16	1.18	.94	.81	1.29	1.63	1.00	1.23	1.53
	Behaviorally General Scale	1.49	1.28	.97	1.39	1.12	1.05	.88	1.44	1.59	1.07	1.39	1.04
Condition 2 t = .34	Behaviorally General Scale	.97	1.10	1.05	1.03	1.06	.91	.92	1.29	1.23	1.29	1.20	1.11
	Behaviorally Specific Scale	1.17	1.12	1.23	1.39	1.12	1.07	.91	.92	1.10	1.10	1.19	1.07
Condition 3 t = .46	Behaviorally General-Mixed Standard Scale	1.31	1.13	1.25	1.45	1.22	1.02	.66	1.52	1.58	.76	1.07	1.52
	Behaviorally Specific Scale	1.23	1.24	1.20	1.37	1.05	1.16	1.23	1.16	1.46	1.13	1.11	1.55

be obtained by comparing the standard deviations of the behaviorally general scales when that scale was presented in the first and second positions. This comparison suggests that an order effect may have occurred since the standard deviations of behaviorally general ratings in the first position were significantly larger than behaviorally general ratings made in the second position. On the basis of the obtained data no firm conclusions about the influence of rating method on the standard deviations of the ratings can be made.

Multitrait-Multirater Analysis

Multitrait-multirater correlation matrices were computed to determine the extent of interrater agreement and the convergent and discriminant validity of each rating method. Each rating condition resulted in two 24 x 24 multitrait (12 performance dimensions) multirater (first and second line supervisor) matrices. These matrices are presented in Tables 12 through 17.

Each matrix is relatively large, making visual inspection according to the Campbell and Fiske (1959) method difficult and judgmental. It is possible, however, to summarize multitrait-multirater correlations in analysis of variance terms. An analysis of variance model developed for this purpose and presented in Kavanaugh, MacKinney and Wolins (1971) provides estimates of the four variance components of the multitrait-multirater situation. These are (a) subject variance, which reflects the overall amount of agreement (convergent validity)

TABLE 12

Intercorrelations Among 12 Dimensions Rated by First and Second Line Supervisors Using Behaviorally General-Mixed Standard Scales: Behaviorally General-Mixed Standard and Behaviorally General Condition

Raters		First Line Supervisors												Second Line Supervisors									
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
First Line	1																								
	2	39																							
	3	54	45																						
	4	53	54	41																					
	5	54	35	42	39																				
	6	22	23	33	30	46																			
	7	36	32	30	30	42	18																		
	8	59	21	33	27	42	25	32																	
	9	40	56	53	44	50	27	33	24																
	10	43	54	23	43	33	20	31	31	23															
	11	46	43	60	22	43	27	52	31	20	29														
	12	67	48	71	49	44	26	30	52	58	37	38													
Second Line	1	(25)	12	31	17	19	42	13	23	15	15	21	35												
	2	20	(18)	20	41	23	32	22	23	17	21	12	29	74											
	3	30	20	(20)	40	23	17	16	29	20	18	01	39	62	72										
	4	25	15	19	(44)	31	32	23	31	11	29	16	25	62	73	63									
	5	24	25	33	19	(44)	39	26	28	33	27	25	41	67	59	57	63								
	6	43	14	17	28	17	(20)	04	24	05	12	15	36	57	50	57	41	42							
	7	21	04	16	17	12	30	(05)	20	01	11	15	20	73	63	59	50	52	57						
	8	42	19	25	20	21	23	05	(57)	08	33	14	33	57	45	48	50	46	53	53					
	9	35	49	46	51	37	22	16	17	(66)	29	18	49	39	54	47	44	42	26	29	31				
	10	29	18	15	32	18	21	03	27	14	(37)	06	34	60	65	67	53	53	58	71	53	33			
	11	29	15	21	08	13	28	00	24	08	29	(23)	32	52	41	38	27	46	58	64	56	24	69		
	12	42	37	41	42	36	32	19	33	38	23	22	(47)	73	72	69	59	58	65	67	56	64	53	45	

Note: Numbers in parentheses indicate validity diagonal; triangles with broken lines equal heterodimension-heterorater triangles; triangles with solid lines equal heterodimension-monorater triangles.

TABLE 13

Intercorrelations Among 12 Dimensions Rated by First and Second Line Supervisors Using Behaviorally General Scales: Behaviorally General-Mixed Standard and Behaviorally General Condition

Raters		First Line Supervisors												Second Line Supervisors											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
First Line	1																								
	2	61																							
	3	50	45																						
	4	45	30	36																					
	5	76	48	27	49																				
	6	54	62	38	42	52																			
	7	41	40	47	35	29	62																		
	8	70	49	53	33	61	45	25																	
	9	36	26	52	45	22	11	12	34																
	10	58	69	44	37	47	49	41	44	25															
	11	42	29	41	27	43	54	61	39	03	34														
	12	59	49	80	43	43	46	46	60	62	49	47													
Second Line	1	(35)	26	54	39	30	21	27	29	50	28	34	61												
	2	34	(46)	38	41	38	28	38	30	40	51	30	48	66											
	3	22	33	(63)	18	15	15	18	28	49	22	18	58	65	39										
	4	32	29	45	(57)	35	27	22	31	56	44	33	50	68	63	38									
	5	39	35	50	40	(38)	30	30	37	54	40	37	58	79	74	52	70								
	6	27	43	52	36	33	(36)	37	33	49	29	35	57	70	67	60	50	69							
	7	50	55	39	49	45	41	(46)	35	30	46	34	52	44	63	23	46	52	41						
	8	46	44	37	19	40	20	10	(58)	21	43	24	42	50	59	46	39	55	43	38					
	9	26	26	41	35	19	12	16	10	(71)	34	10	54	49	35	48	59	49	47	28	20				
	10	49	56	55	38	41	30	42	45	47	(56)	24	63	54	73	43	50	65	61	65	56	38			
	11	39	37	46	48	39	37	39	29	43	46	(48)	54	69	79	33	70	70	61	62	46	39	60		
	12	35	44	64	38	25	21	20	36	60	48	21	(71)	72	65	73	60	72	66	51	59	59	68	63	

Note: Numbers in parentheses indicate validity diagonal; triangles with broken lines equal heterodimension-heterorater triangles; triangles with solid lines equal heterodimension-monorater triangles.

TABLE 14

Intercorrelations Among 12 Dimensions Rated by First and Second Line Supervisors Using Behaviorally General Scales: Behaviorally General and Behaviorally Specific Condition

Raters		First Line Supervisors												Second Line Supervisors											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
First Line	1																								
	2	71																							
	3	77	69																						
	4	65	50	49																					
	5	79	69	65	71																				
	6	68	52	60	57	57																			
	7	54	31	34	58	65	42																		
	8	59	57	70	52	46	60	21																	
	9	66	69	65	47	54	44	36	43																
	10	56	75	53	43	43	38	25	43	62															
	11	67	49	54	47	63	57	76	34	46	44														
	12	75	76	88	45	64	53	36	66	74	62	52													
Second Line	1	(41)	47	50	16	28	24	-04	30	50	35	16	52												
	2	61	(68)	60	33	46	35	14	49	65	49	34	70	65											
	3	56	51	(54)	50	38	48	15	54	45	39	28	54	61	70										
	4	53	61	57	(32)	43	46	05	44	63	39	27	53	76	63	43									
	5	61	58	64	31	(59)	43	27	39	59	33	41	61	79	68	56	77								
	6	45	48	44	45	31	(45)	33	46	53	42	36	52	40	60	67	38	44							
	7	46	30	35	51	44	30	(42)	43	27	21	29	41	36	53	72	22	42	60						
	8	64	50	63	40	39	45	13	(67)	49	35	25	64	55	76	77	49	58	59	66					
	9	56	53	45	42	40	39	27	48	(59)	38	28	52	46	67	73	48	59	59	55	61				
	10	43	51	39	33	32	26	03	39	47	(42)	17	50	57	79	71	45	49	60	55	64	56			
	11	64	67	66	46	56	50	27	55	62	43	(47)	67	67	81	67	68	74	53	45	66	68	61		
	12	64	67	68	48	51	43	30	53	66	51	42	(74)	68	82	75	64	72	67	59	70	72	65	79	

Note: Numbers in parentheses indicate validity diagonal; triangles with broken lines equal heterodimension-heterorater triangles; triangles with solid lines equal heterodimension-monorater triangles.

Table 15

Intercorrelations Among 12 Dimensions Rated by First and Second Line Supervisors Using Behaviorally Specific Scales: Behaviorally General and Behaviorally Specific Condition

Raters		First Line Supervisors												Second Line Supervisors											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
First Line	1																								
	2	56																							
	3	63	38																						
	4	72	52	62																					
	5	78	67	62	70																				
	6	70	38	56	56	72																			
	7	52	37	62	61	52	42																		
	8	74	29	59	55	67	56	38																	
	9	53	63	67	61	53	44	59	33																
	10	57	67	30	61	60	42	53	34	53															
	11	48	43	62	58	55	54	58	36	51	46														
	12	80	74	60	75	79	64	54	58	68	65	56													
Second Line	1	(69)	60	56	63	69	55	58	50	61	42	47	75												
	2	61	(64)	46	56	68	40	37	41	48	47	46	73	76											
	3	72	62	(63)	64	76	62	43	50	54	37	50	77	81	84										
	4	38	46	38	(54)	58	31	43	19	53	32	40	60	73	67	64									
	5	58	48	49	59	(54)	44	63	36	61	44	43	69	79	66	64	59								
	6	68	49	58	63	63	(55)	52	52	58	35	41	72	82	73	84	65	79							
	7	56	42	41	48	52	54	(39)	36	50	35	30	63	68	63	68	55	72	70						
	8	61	31	54	46	56	50	27	(63)	29	18	29	62	65	53	65	41	39	58	46					
	9	50	53	46	53	68	38	40	34	(46)	37	35	72	69	78	75	65	65	66	63	45				
	10	55	55	37	44	55	38	34	35	45	(32)	34	66	81	73	71	66	70	72	65	46	70			
	11	64	57	60	57	70	59	42	51	55	35	(43)	74	75	69	80	60	72	82	73	62	73	65		
	12	62	57	48	56	64	42	46	45	50	43	34	(75)	84	85	80	69	76	77	69	61	79	78	77	

Note: Numbers in parentheses indicate validity diagonal; triangles with broken lines equal heterodimension-heterorater triangles; triangles with solid lines equal heterodimension-monorater triangles.

TABLE 16

Intercorrelations Among 12 Dimensions Rated by First and Second Line Supervisors Using Behaviorally General-Mixed Standard Scales: Behaviorally General-Mixed Standard and Behaviorally Specific Condition

Raters		First Line Supervisors												Second Line Supervisors									
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
First Line	1																								
	2	10																							
	3	70	15																						
	4	07	55	03																					
	5	22	69	23	65																				
	6	42	-13	24	09	10																			
	7	22	48	00	74	58	-06																		
	8	74	-04	54	-21	07	26	-25																	
	9	65	14	46	-17	13	29	-39	63																
	10	00	62	03	50	60	-01	47	-02	-05															
	11	39	66	42	34	70	12	38	24	25	37														
	12	30	65	28	64	68	20	51	26	27	51	50													
Second Line	1	(65)	06	50	-14	27	17	-23	52	51	-03	33	15												
	2	54	(16)	34	-15	04	25	-39	53	56	-11	00	00	72											
	3	48	-02	(52)	00	16	18	-22	43	51	04	08	10	76	78										
	4	53	-06	41	(09)	19	21	-31	40	48	02	13	10	84	80	84									
	5	46	29	40	19	(51)	25	01	26	40	25	41	41	77	58	70	82								
	6	43	-00	39	00	11	(12)	24	46	52	-08	18	09	50	64	50	46	36							
	7	36	-23	26	01	-09	27	(-09)	28	38	-23	-17	02	31	58	41	32	21	44						
	8	62	25	53	16	44	30	01	(62)	50	23	45	41	68	49	52	56	62	49	29					
	9	54	18	45	-14	10	14	-33	43	(61)	-06	29	19	70	50	58	57	53	45	29	51				
	10	42	00	36	04	15	15	-07	32	26	(14)	04	15	54	59	57	50	42	33	59	46	42			
	11	44	-01	36	09	16	15	02	24	30	-13	(12)	13	44	35	32	25	23	43	61	42	34	50		
	12	70	10	51	-03	29	23	-23	51	54	07	29	(24)	66	69	69	77	72	55	41	70	70	67	47	

Note: Numbers in parentheses indicate validity diagonal; triangles with broken lines equal heterodimension-heterorater triangles; triangles with solid lines equal heterodimension-monorater triangles.

TABLE 17

Intercorrelations Among 12 Dimensions Rated by First and Second Line Supervisors Using Behaviorally Specific Scales: Behaviorally General-Mixed Standard and Behaviorally Specific Condition

Raters		First Line Supervisors												Second Line Supervisors											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
First Line	1																								
	2	37																							
	3	60	36																						
	4	66	48	63																					
	5	86	38	68	67																				
	6	51	29	46	36	51																			
	7	68	27	46	40	54	46																		
	8	63	43	52	60	53	56	39																	
	9	72	37	51	52	69	45	57	44																
	10	56	46	43	34	59	44	54	40	49															
	11	58	59	46	42	61	59	53	35	61	57														
	12	78	54	68	69	74	55	55	71	74	57	63													
Second Line	1	(54)	38	52	52	58	34	40	56	56	29	43	64												
	2	51	(39)	52	58	49	37	41	63	58	29	42	62	76											
	3	52	22	(42)	54	51	33	45	46	55	17	31	51	76	71										
	4	61	40	60	(62)	60	30	37	58	51	35	42	67	77	80	66									
	5	31	18	32	34	(37)	44	29	50	43	19	27	40	70	70	72	53								
	6	46	24	30	39	45	(42)	42	45	50	27	34	41	66	64	86	58	70							
	7	52	29	48	48	54	38	(45)	50	64	24	42	64	75	71	85	67	67	75						
	8	49	23	46	50	46	33	35	(64)	45	20	30	56	83	79	71	72	70	61	68					
	9	50	35	50	59	52	24	33	54	(58)	16	30	56	76	83	79	77	66	68	80	72				
	10	55	39	57	51	54	40	41	54	59	(31)	41	59	86	80	68	75	72	65	71	80	80			
	11	53	06	30	38	49	30	46	45	53	23	(30)	44	61	58	77	56	67	74	74	62	63	50		
	12	52	32	52	55	54	31	36	52	61	16	36	(61)	82	85	82	76	71	70	85	80	89	82	66	

Note: Numbers in parentheses indicate validity diagonal; triangles with broken lines equal heterodimension-heterorater triangles; triangles with solid lines equal heterodimension-monorater triangles.

on subjects over sources (raters) and traits (dimensions); (b) subject by trait variance, which indicates the degree of rater discriminations on traits by subjects (discriminant validity); (c) subject by source variance, which indicates the amount of source bias (halo) in the ratings; and, (d) error variance.

The estimation of these four sources of variance provides evidence interpretable for within-matrix comparisons, but not across matrices. Since it is also of interest to compare across matrices relative to convergent and discriminant validity, Kavanaugh, MacKinney and Wolins derived a comparison index using the formula:

$$\text{comparison index} = \frac{\text{true variance}}{\text{true variance plus error variance}}$$

When applied to each variance component, this index indicates the amount of convergence, discrimination and method bias in a form amenable to intermatrix comparisons. The usefulness of this index in the present study is that the comparison index permits comparison of convergent validity, discriminant validity and halo across different rating methods.

The multitrait-multirater matrices for the mixed standard and behaviorally general scale comparison of experimental condition 1 are provided in Tables 12 and 13 respectively. Table 18 presents the analysis of variance data and comparison indexes obtained from these two matrices. Since the validity diagonals in each matrix represent a critical characteristic of performance evaluations, i.e., interrater reliability, the

TABLE 18

Analysis of Variance for Multitrait-Multirater Correlation Matrices

Experimental Condition		Source of Variance	df	MS	F	Variance Component[1]	Comparison Index
Condition 1	Behaviorally General-Mixed Standard Scale	Subject	59	9.29	21.06**	.37	.46
		Subject x Trait	660	.64	1.49*	.10	.18
		Subject x Source	59	3.24	7.36**	.23	.35
		Error	660	.44		.44	
	Behaviorally General Scale	Subject	59	11.35	32.14**	.46	.57
		Subject x Trait	660	.66	1.86*	.11	.24
		Subject x Source	59	1.95	5.52**	.13	.27
		Error	660	.35		.35	
Condition 2	Behaviorally General Scale	Subject	52	13.12	39.92**	.35	.62
		Subject x Trait	572	.50	1.53*	.08	.21
		Subject x Source	52	2.20	6.70**	.16	.32
		Error	572	.33		.33	
	Behaviorally Specific Scale	Subject	52	14.36	42.70**	.58	.64
		Subject x Trait	572	.42	1.24*	.04	.11
		Subject x Source	52	1.83	5.45**	.13	.27
		Error	572	.34		.34	
Condition 3	Behaviorally General-Mixed Standard Scale	Subject	63	8.19	15.67**	.32	.38
		Subject x Trait	693	.66	1.26*	.07	.12
		Subject x Source	63	3.20	6.11	.22	.30
		Error	693	.52		.52	
	Behaviorally Specific Scale	Subject	63	3.46	40.08**	.55	.62
		Subject x Trait	693	.41	1.21*	.04	.10
		Subject x Source	63	2.75	8.19**	.20	.38
		Error	693	.34		.34	

[1]Estimates for Variance Components were computed from Kavanaugh, MacKinney, and Wolins (1971) as follows:

$$\text{VC subject} = \frac{\text{MS subject} - \text{MS error}}{nm}$$

$$\text{VC subject x trait} = \frac{\text{MS subject x trait} - \text{MS error}}{m}$$

$$\text{VC subject x source} = \frac{\text{MS subject x source} - \text{MS error}}{n}$$

* $p < .01$
** $p < .005$

differences in these diagonals were tested for statistical significance. The correlations in each diagonal (in parentheses) were first converted to Z scores. Difference scores were then obtained and used to compute t-tests of the difference between the diagonals for the two rating methods. The significance level reached .01 (t = 4.38) in favor of the obviously larger interrater reliabilities of the behaviorally general method. Subject variance[1] (the degree of agreement among raters on subject by tracts), as indicated by the comparison index (Table 18), was also greater for the behaviorally general method, reflecting the higher overall magnitude of the correlations in the behaviorally general matrix. Evidence of discriminant validity[2] requires that, for a given dimension, the value in the validity diagonal must be greater than the values in its column and row in the heterodimension-heterorater and heterodimension-monrater triangles. The degree to which this requirement is met for the two methods is difficult to assess by visual inspection of Tables 12 and 13. However, from the Subject by Trait comparison indices[2] it can be seen that discriminant validity is greater in the behaviorally

---

[1]The sum of squares for convergent validity is computed as follows: SS Subjects = Nnm ($\bar{r}o$), where $\bar{r}o$ = average correlation of all elements in the matrix, including the ones in the main diagonals; N = number of subjects; n = number of traits; m = number of sources.

[2]The computation for discriminant validity is: SS Subject x Trait = Nnm ($\bar{r}wt$ - $\bar{r}o$), where $\bar{r}wt$ = average correlation between sources within traits; computation - the sum of the validity diagonal times two plus nm divided by nm^2.

general ratings. The degree to which the ratings for the two methods exhibited halo bias is also not obvious from looking at the matrices. Based upon the comparison indices[3], however, the behaviorally general method was less biased. It was concluded that the behaviorally general ratings were clearly superior to the mixed standard scale ratings in terms of multitrait-multirater analyses. Interrater reliability, Subject Variance, and Subject by Trait Variance were all greater while halo variance was less when raters used the behaviorally general method.

Tables 14 and 15 contain the multitrait-multirater matrices for the behaviorally general and behaviorally specific scale comparison of experimental condition 2. The validity diagonals for these two obvious scale continuum methods appear to be about the same magnitude. A t-test revealed that there was no statistical difference in the interrater reliabilities of the two methods ($t = .66$). The Subject Variance comparison indexes (Table 18) show that the degree of agreement among raters on Subjects by Traits was also about equal for the two methods. The comparison indices also reveal that the behaviorally general rating method resulted in greater discriminant validity while halo bias was less with the behaviorally specific method. It was therefore concluded that the multitrait-multirater results were a standoff between these two rating scale

---

[3]Halo is computed as: SS Subject x Source = Nnm ($\bar{r}ws - \bar{r}o$), where $\bar{r}ws$ = average correlation between traits within sources; computation - the sum of the heterodimension-monorater triangles times two plus nm divided by mn^2. Computation of the Error term is: SS Error = Nnm ($1 - \bar{r}wt - \bar{r}ws + \bar{r}o$).

methods: no differences were found in interrater reliability and Subject Variance, while the behaviorally general methods had greater Subject by Trait Variance, and the behaviorally specific method has less halo.

The multitrait-multirater analyses for experimental condition 3, mixed standard scale compared with behaviorally specific scale, are presented in Tables 16 and 17. The validity diagonals are obviously greater in the behaviorally specific matrix. A t-test confirmed that the difference was significant at the .05 level ($t = 2.30$). Subject Variance, illustrated in Table 18 by the comparison indices, was also greater for the behaviorally specific ratings. The comparison indices for discriminant validity are essentially equal for the two methods. These indices are close, despite the lower validity diagonal in the mixed standard scale matrix, because the comparison index reflects the lower overall correlations in this matrix. The comparison indices show that halo bias was slightly lower in the mixed standard scale ratings. The reason for the higher behaviorally specific halo index is that trait intercorrelations were very high for second line raters. Overall then, the multitrait-multirater results favor the behaviorally specific method. There was no difference between the two methods in terms of discriminant validity and halo bias was slightly less in the mixed standard scale ratings. However, interrater reliability and Subject Variance were both much larger in the behaviorally specific ratings.

Briefly summarizing the results of the multitrait-multirater

analyses for all three experimental conditions, interrater reliability was significantly larger for the scales employing an obvious scale continuum. Whether the scale anchors were behaviorally general or specific did not affect the magnitude of the validity diagonals, however, so long as the scale continuum was obvious. The overall magnitude of the correlations in an entire matrix were also found to be greater where the scale continuum of the rating method was obvious. Thus, the Subject Variance (degree of agreement among raters on Subjects by Traits) was higher for both the behaviorally general and behaviorally specific methods in comparison to the mixed standard scale. The use of behaviorally general anchors on an obvious continuum scale (the behaviorally general rating method) resulted in higher discriminant validity than when the behaviorally general anchors were mixed or when behaviorally specific anchors were presented on an obvious continuum. Significant halo bias was present in every rating method. Furthermore, the results did not consistently favor any one rating scale as a method of reducing halo bias. In conclusion, the multitrait-multirater results favored both obvious scale continuum methods over the mixed standard scale while there appeared to be little difference in the results when these two obvious scale methods were compared to one another.

The Reproducibility Coefficient as an Estimate of Rater Reliability

Blanz and Ghiselli (1972) propose a supplement to the usual parallel raters or rating-rerating methods of determining

the reliability of scales. They contend that reproducibility coefficients calculated with the mixed standard scale format may be a suitable alternative in estimating reliability. If such is the case, reproducibility coefficients and reliability coefficients for the same rating scales should correlate highly. To examine this contention, correlations between the 12 reproducibility coefficients obtained by both the Good-enough and "logic" methods and the 12 interrater reliability coefficients were compared for the mixed standard scales in experimental conditions 1 and 3. The resulting correlations (Table 19) were low and not statistically significant. The use of reproducibility coefficients as a viable alternative to the more traditional methods of estimating reliability is not supported by these data.

Factor Analyses

Halo effects for the three rating methods were further analyzed by factor analysis, utilizing the principal factor solution with varimax rotation (Harman, 1970). Initial computer runs were directed toward determining the number of factors which could most appropriately be specified across all factor solutions. Only the ratings by first line supervisors were employed in this step since they were judged to be superior to the second line ratings. Each principal factor solution was rotated using in succession a two, three, and four factor solution. In most instances, the four factor solution did not produce loadings of significant magnitude in

TABLE 19

Correlation[1] of the Interrater Reliabilities and Reproducibility Coefficients of Mixed, Standard Scales

Comparison	
Goodenough Reproducibility Coefficients with Interrater Reliabilities	
Experimental Condition 1	-.06
Experimental Condition 3	-.36
"Logic" Reproducibility Coefficients with Interrater Reliabilities	
Experimental Condition 1	.23
Experimental Condition 2	.08

[1]Pearson Product-moment coefficient

the fourth factor. It was clear, however, that more than two factors were needed for all but one set of ratings. Therefore, the three factor solution was used to compare the factors obtained using the different rating methods.

The rotated factor matrices for the first line ratings in each experimental condition are displayed in Tables 20 through 22. To facilitate inspection of the matrices, factor loadings of .30 or less have been omitted. With the exception of the mixed standard scale ratings in experimental condition 3 (Table 22), factor analysis of each set of ratings yielded three clearly distinguishable factors. In addition, percent of total factor variance was relatively evenly spread in each factor pattern. The best solution for the mixed standard scale ratings in condition 3, in terms of factor clarity, was obtained when a two factor pattern was specified. Thus, in this condition, behaviorally specific ratings were superior to the mixed standard scale ratings in terms of the number of factors. In the remaining two experimental conditions, however, there was not a consistent difference between methods in factor definition.

For purposes of comparison, the ratings of the second line supervisors were also factor analyzed using a three factor solution. The results of these analyses are presented in Tables 23 through 25. As these tables show, the factor patterns for second line ratings were not as clear as those obtained for first line ratings. Furthermore, specifying a two factor solution did not improve clarity of factor definition

TABLE 20

Factor Analysis[1]: Rotated Factor Matrices,[2]

First Line Ratings in Experimental Condition 1

	Performance Dimension	Behaviorally General-Mixed Standard Scale			Behaviorally General Scale		
		I	II	III	I	II	III
1.	Planning & Organizing	.70			.32	.78	
2.	Understanding & Carrying out Instructions		.73			.61	.35
3.	Work Effort	.60	.31	.35	.72		.40
4.	Selecting New Employees	.32	.60		.36	.41	
5.	Effective Use of Employees	.45		.45		.81	
6.	Rewarding & Motivating Employees					.50	.63
7.	Customer Relations			.53			.76
8.	Store Cleanliness	.57			.38	.63	
9.	Control of Shrinkage	.43	.59		.75		
10.	Accuracy of Work		.49	.32		.55	.35
11.	Sales			.75			.68
12.	Overall Effectiveness	.73	.41		.76		
	Percent of Total Factor Variance	39.3	33.5	27.1	29.9	40.8	29.3

[1]Principal Factor Solution with Varimax Rotation

[2]Factor loadings of .30 or smaller omitted

TABLE 21

Factor Analysis[1]: Rotated Factor Matrices,[2]

First Line Ratings in Experimental Condition 2

	Performance Dimension	Behaviorally General Scale I	Behaviorally General Scale II	Behaviorally General Scale III	Behaviorally Specific Scale I	Behaviorally Specific Scale I	Behaviorally Specific Scale III
1.	Planning & Organizing	.52	.51	.53	.76	.41	
2.	Understanding & Carrying out Instructions	.36	.78			.82	
3.	Work Effort	.65	.56		.48		.74
4.	Selecting New Employees	.46		.58	.49	.44	.51
5.	Effective Use of Employees	.39	.39	.67	.69	.51	.31
6.	Rewarding & Motivating Employees	.59		.43	.67		.33
7.	Customer Relations			.87			.67
8.	Store Cleanliness	.74	.34		.78		
9.	Control of Shrinkage		.70			.50	.64
10.	Accuracy of Work		.75			.71	
11.	Sales		.33	.76			.61
12.	Overall Effectiveness	.52	.72		.54	.63	.38
	Percent of Total Factor Variance	28.9	37.5	33.6	37.0	31.8	31.3

[1]Principal Factor Solution with Varimax Rotation

[2]Factor loadings of .30 or smaller omitted

TABLE 22

Factor Analysis[1]: Rotated Factor Matrices,[2]

First Line Ratings in Experimental Condition 3

	Performance Dimension	Behaviorally General-Mixed Standard Scale I	Behaviorally General-Mixed Standard Scale II	Behaviorally General-Mixed Standard Scale III	Behaviorally Specific Scale I	Behaviorally Specific Scale II	Behaviorally Specific Scale III
1.	Planning & Organizing	.89			.59		.71
2.	Understanding & Carrying out Instructions		.84	.34	.32	.67	
3.	Work Effort	.68			.60		.40
4.	Selecting New Employees		.79	.32	.73		
5.	Effective Use of Employees		.86		.58		.67
6.	Rewarding & Motivating Employees	.37		.44	.31	.36	.45
7.	Customer Relations	.34	.75				.69
8.	Store Cleanliness	.81			.70		
9.	Control of Shrinkage	.79			.39		.65
10.	Accuracy of Work		.68			.42	.55
11.	Sales	.39	.67			.63	.57
12.	Overall Effectiveness		.77		.66	.40	.49
	Percent of Total Factor Variance	39.6	53.0	7.4	38.2	20.9	40.9

[1]Principal Factor Solution with Varimax Rotation

[2]Factor loadings of .30 or smaller omitted

TABLE 23

Factor Analysis[1]: Rotated Factor Matrices,[2]

Second Line Ratings in Experimental Condition 1

Performance Dimension	Behaviorally General-Mixed Standard Scale			Behaviorally General Scale		
	I	II	III	I	II	III
1. Planning & Organizing	.33	.64	.49	.51	.57	.43
2. Understanding & Carrying out Instructions	.43	.72	.31		.33	.81
3. Work Effort	.45	.55	.40	.80		
4. Selecting New Employees	.34	.72			.70	.44
5. Effective Use of Employees		.57	.39	.41	.52	.58
6. Rewarding & Motivating Employees	.39		.65	.52	.39	.46
7. Customer Relations		.44	.68			.69
8. Store Cleanliness			.58	.45		.53
9. Control of Shrinkage	.61	.32		.36	.61	
10. Accuracy of Work		.47	.71	.38		.72
11. Sales			.81		.51	.71
12. Overall Effectiveness	.71	.40	.43	.67	.40	.46
Percent of Total Factor Variance	23.4	36.0	40.6	29.0	27.7	43.4

[1]Principal Factor Solution with Varimax Rotation

[2]Factor loadings of .30 or smaller omitted

TABLE 24

Factor Analysis[1]: Rotated Factor Matrices,[2]
Second Line Ratings in Experimental Condition 2

	Performance Dimension	Behaviorally General Scale I	Behaviorally General Scale II	Behaviorally General Scale III	Behaviorally Specific Scale I	Behaviorally Specific Scale II	Behaviorally Specific Scale III
1.	Planning & Organizing			.79	.49	.52	.58
2.	Understanding & Carrying out Instructions	.53	.55	.55	.41	.74	.34
3.	Work Effort	.82		.34	.65	.56	.37
4.	Selecting New Employees			.84		.59	.42
5.	Effective Use of Employees	.34		.83		.40	.79
6.	Rewarding & Motivating Employees	.68			.53	.39	.64
7.	Customer Relations	.79			.35	.39	.61
8.	Store Cleanliness	.69		.38	.71		
9.	Control of Shrinkage	.67			.33	.69	.38
10.	Accuracy of Work	.55	.56	.32		.60	.52
11.	Sales	.47		.68	.58	.39	.56
12.	Overall Effectiveness	.63		.60	.44	.67	.47
	Percent of Total Factor Variance	46.0	11.9	42.2	28.0	37.6	34.4

[1]Principal Factor Solution with Varimax Rotation

[2]Factor loadings of .30 or smaller omitted

TABLE 25

Factor Analysis[1]: Rotated Factor Matrices,[2]

Second Line Ratings in Experimental Condition 3

Performance Dimension	Behaviorally General-Mixed Standard Scale			Behaviorally Specific Scale		
	I	II	III	I	II	III
1. Planning & Organizing	.80	.42		.36	.70	.44
2. Understanding & Carrying out Instructions	.33	.74	.45	.41	.75	.33
3. Work Effort	.50	.68		.32	.44	.77
4. Selecting New Employees	.61	.73			.80	.39
5. Effective Use of Employees	.72	.47		.60	.37	.50
6. Rewarding & Motivating Employees	.34	.32	.47		.33	.79
7. Customer Relations			.79	.31	.51	.67
8. Store Cleanliness	.68		.31	.39	.66	.41
9. Control of Shrinkage	.63				.71	.50
10. Accuracy of Work	.34	.31	.59	.45	.74	.31
11. Sales			.72			.78
12. Overall Effectiveness	.77	.34	.40	.51	.62	.36
Percent of Total Factor Variance	48.3	28.4	28.3	17.1	37.7	45.2

[1]Principal Factor Solution with Varimax Rotation

[2]Factor loadings of .30 or smaller omitted

for second line ratings. It was concluded that the factor solution of the ratings of second line supervisors was not very well defined and that second line ratings were not sensitive to differences in methods.

Inspection of the factor patterns of first line ratings shows that the twelve performance dimensions tended to load on similar factors, regardless of the rating method being used. The first factor typically consisted of the dimensions: Planning and Organizing, Work Effort, Store Cleanliness, and Overall Effectiveness. These dimensions appear to have in common a diligence or conscientiousness factor. The second factor's highest loadings were: Understanding and Carrying Out Instructions, Selecting New Employees, and Accuracy of Work. These performance dimensions imply a common intellectual factor. Control of Shrinkage loaded equally on both the first and second factors indicating that this dimension contains both a diligence and an intellectual component. The third factor appears to represent skill in dealing with people. This factor contained the highest loadings for: Rewarding and Motivating Employees, Customer Relations, and Sales. The only performance dimension that did not load with some degree of consistency on one of those factors was Effective Use of Employees.

In order to further examine the degree of similarity among factors obtained using the various rating methods, coefficients of congruence (Tucker, 1951) were computed between all possible pairs of factors for each rating method. The coefficient of

congruence, which ranges in value from +1 to -1, should not be interpreted in the same manner as a correlation coefficient. This index is highly influenced by the level and sign of the factor loadings. Even if the patterns are unrelated, factors with high loadings will tend to have a high coefficient of congruence. In the present study, congruence coefficients as high as .70 were obtained in situations where close scrutiny of the factor loadings revealed that there was no relationship between the factors. The relative magnitudes of the coefficients within a given congruence matrix are therefore a basis for objective comparison in this study, but the basic factor patterns provide the primary source of information concerning factor similarity. The coefficient of congruence is admittedly a crude measure, but may be used as a proportionality criterion to supplement to the more usual alternative of rough data inspection and personal impressions.

When using the coefficient of congruence, it is generally recommended that each factor in one study or method be compared with all the factors of the second and be paired to the one factor with which it has the highest coefficient (Harman, 1970, pp. 271). In the present study, coefficients of congruence were computed to make two separate comparisons of the factor loadings. First, factor patterns obtained by the same rating method were compared. This comparison reveals the similarity of factor patterns based upon the rating method. Second, factor patterns obtained by the same group of first line raters using two rating methods were compared. This

comparison reveals the extent to which ratings performed by the same raters provide similar factor patterns despite the use of different rating methods.

Coefficients of congruence for the first comparison, that of different raters using the same rating method, are presented in Table 26. The highest coefficient in each column and row has been underlined to indicate the strongest pairings. Since factor I in each set of ratings is the diligence factor, factor II, the intellectual factor, and factor III, the people skill factor; the ideal pairings would be for each factor to pair with its corresponding number in the other set. Only the two behaviorally specific scales paired in this manner. The behaviorally general scales, while pairing on three factors, did not have all pairings in the main diagonal of the congruence matrix. In the mixed standard scale there was not a pairing for factor III of the third experimental condition rating set. This was further evidence that only two factors were descriptive of this set of ratings.

The magnitudes of the paired coefficients in the behaviorally specific matrix were the highest of the three matrices, while those in the mixed standard scale matrix were the smallest. Although the magnitude of a congruence coefficient is highly influenced by several numerical features which are unrelated to the actual congruence of factor loadings, the data presented in Tables 20, 21, and 22 appear to support a conclusion that factors obtained when raters used the behaviorally specific scales were more similar than factors

TABLE 26

Coefficients of Congruence Comparing the Factors Obtained by Different Raters Using Identical Rating Methods[1]

Rating Method			Coefficient of Congruence		
			Experimental Condition 2		
			I	II	III
Behaviorally General Scale	Experimental Condition 1	I	.83	.87	.57
		II	.87	.83	.77
		III	.69	.66	.87
			Experimental Condition 3		
			I	II	III
Behaviorally General-Mixed Standard Scale	Experimental Condition 1	I	.80	.58	.38
		II	.45	.80	.12
		III	.49	.78	.19
			Experimental Condition 3		
			I	II	III
Behaviorally Specific Scale	Experimental Condition 2	I	.95	.72	.83
		II	.78	.88	.82
		III	.82	.78	.91

[1]First line ratings only

obtained when raters used either of the other two rating methods.

Coefficients of congruence for comparison of the same raters using two different rating methods are presented in Table 27. In experimental condition 2, comparing the rating methods with obvious scale continuums, the coefficient of congruent pairings were in the main diagonal of the matrix. In both of the remaining matrices the mixed standard scale failed to produce a pairing for one of the three factors. In fact, the pairings for the mixed standard scale in experimental condition 3 came very close to representing only one factor. Finally, the coefficients were highest in the second experimental condition where both rating methods employed an obvious scale continuum.

In summary, evidence based upon the coefficient of congruence statistic indicates that the method used for rating did make a difference in the performance factors which were obtained. This evidence favored the factor superiority of the behaviorally specific scale, especially over the mixed standard scale rating method. The pairings in the congruence matrix were completely consistent only for the behaviorally specific rating method. Furthermore, the results were consistently negative for the mixed standard scale rating method. Regardless of whether the comparison involved the same rating method with different raters or the same raters with different rating methods, the mixed standard scale was the only method which failed to produce a pairing for all three factors.

TABLE 27

Coefficients of Congruence Comparing the Factors Obtained by the Same Raters Using Different Rating Methods[1]

Experimental Condition			Coefficient of Congruence		
			Behaviorally General Scale		
			I	II	III
Condition 1	Behaviorally General-Mixed Standard Scale	I	.90	.83	.66
		II	.80	.79	.64
		III	.57	.79	.92
			Behaviorally Specific Scale		
			I	II	III
Condition 2	Behaviorally General Scale	I	.96	.72	.79
		II	.74	.93	.80
		III	.78	.74	.88
			Behaviorally Specific Scale		
			I	II	III
Condition 3	Behaviorally General-Mixed Standard Scale	I	.68	.51	.61
		II	.67	.80	.72
		III	.44	.03	.34

[1]First line ratings only

Rating Method - Test Score Correlations

Where performance evaluations are used in test validation, an important consideration in judging the utility of any given rating device is the extent to which performance evaluations are predictable from appropriate test scores. Before comparing the correlations between the test scores and the various rating methods in the present study, it was first necessary to postulate relationships between test scales and various performance dimensions. The basis for assumption of a relationship was a comparison of the psychological construct as defined for a given scale in the test manual with the definition of a performance dimension and the factor analysis results for that dimension. Using this decision process, various scales from the test battery were hypothesized to be related (either positively or negatively) to eight of the twelve performance dimensions. In some cases a test scale was hypothesized to correspond to more than one performance dimension.

The correlations between these preselected performance dimensions and predictor scores are presented for first line ratings in Table 28 and second line ratings in Table 29. Since the validity coefficients in this study tended to be low, a lenient significance level of .10 was chosen to identify significant relationships between predictor and criteria. In addition, within each experimental condition, a count was made of the number of validity coefficients that were higher (in the proper direction) for one rating method over another.

TABLE 28

Correlations[@] of Selected Predictors and First Line Ratings

		First Line Raters					
		Condition 1 (N=56)		Condition 2 (N=50)		Condition 3 (N=63)	
Performance Rating	Predictor	Mixed Standard Scale	General Scale	General Scale	Specific Scale	Mixed Standard Scale	Specific Scale
Planning & Organizing	Orderliness	.19	.08	.15	.10	-.03	-.14
	Social°	-.12	-.05	.12	.00	-.21*	-.08
	Artistic°	-.25*	-.17	-.04	-.19	-.26**	-.35***
	Impulsive°	-.34**	-.14	-.06	.02	-.10	-.16
Work Effort	Active	.10	.12	.31**	.25**	.22*	.12
Store Cleanliness & Appearance	Tough Minded°	.04	.04	.23	.17	-.17	-.23*
Understanding & Carrying Out Instructions	Adaptability Test	.08	.14	.22	.21	.21	.24*
	Achievement	.05	.18	.21	.39***	.31**	.43***
Effective Use of Employees	Impulsive°	-.21	-.15	.05	-.06	-.22*	-.24*
Accuracy of Work	Adaptability Test	.11	.12	.34**	.30**	.15	.15
	Variety°	-.07	-.26*	-.28**	-.02	-.09	-.16
	Conventional	.18	.20	.23	.27*	.11	.10
Sales	Enterprising	.10	.11	.10	.14	.17	.01
	Ascendant	-.04	.21	.03	.19	.07	.15
Overall Effectiveness	Achievement	-.24*	-.07	.24*	.24*	.17	.28**
	Intellectual°	-.09	-.06	-.02	-.05	-.09	-.30**
	Artistic°	-.24*	-.17	-.09	-.19	-.04	-.34***
	Infrequency	.17	.10	-.07	.06	.14	.24*
	Active	.23*	.18	.19	.30**	.14	.18
Number of significant correlations		5	1	4	6	5	9
Number of higher magnitude correlations		9	9	6	12	5	12

[@]Bivariate Subsample Method for Missing Data

° Inverse Relationship Hypothesized

* $p < .10$

** $p < .05$

*** $p < .01$

TABLE 29

Correlations[@] of Selected Predictors and Second Line Ratings

		Second Line Raters					
		Condition 1 (N=56)		Condition 2 (N=50)		Condition 3 (N=63)	
Performance Rating	Predictor	Mixed Standard Scale	General Scale	General Scale	Specific Scale	Mixed Standard Scale	Specific Scale
Planning & Organizing	Orderliness	.12	.11	-.08	-.01	-.13	-.09
	Social°	-.14	-.33**	.18	.08	-.28**	-.35***
	Artistic°	-.14	-.17	.04	.02	-.27**	-.32**
	Impulsive°	-.13	-.10	-.09	.06	-.18	-.09
Work Effort	Active	.27**	.38***	.20	.23	.15	.21
Store Cleanliness & Appearance	Tough Minded°	.12	-.01	-.06	-.07	-.12	.00
Understanding & Carrying Out Instructions	Adaptability Test	-.04	.09	.07	-.03	-.13	-.03
	Achievement	.08	.06	.03	.08	.18	.34***
Effective Use of Employees	Impulsive°	-.09	-.12	-.07	.04	-.16	-.07
Accuracy of Work	Adaptability Test	-.07	.00	.11	-.10	-.24	-.19
	Variety°	-.14	-.18	.12	.01	-.13	-.26**
	Conventional	.08	.16	.03	-.07	-.15	-.12
Sales	Enterprising	-.05	.03	.09	.19	.02	-.19
	Ascendant	.17	.05	.02	.00	.19	.26*
Overall Effectiveness	Achievement	-.11	-.02	.09	.15	.32**	.31**
	Intellectual°	-.01	.05	.00	.04	-.24*	-.17
	Artistic°	-.19	-.12	.00	-.02	-.29**	-.32**
	Infrequency	.02	.04	-.18	-.16	.08	.09
	Active	.24*	.18	.29**	.24*	.14	.14
Number of significant correlations		2	2	1	1	5	7
Number of higher magnitude correlations		6	12	7	11	6	12

[@]Bivariate Subsample Method for Missing Data ° Inverse Relationship Hypothesized

* $p < .10$
** $p < .05$
*** $p < .01$

As might be expected, both the overall magnitudes of the correlations and the number of correlations achieving significance were greater when ratings performed by first line supervisors were correlated with the tests than when the ratings by second line raters were used. Looking first at the data for first line raters (Table 28), it appears that the behaviorally specific ratings were the most predictable. In comparison to both the behaviorally general scale (experimental condition 2) and the mixed standard scale (experimental condition 3), the behaviorally specific scale showed the greater number of significant correlations and the greater number of higher magnitude correlations. When the mixed standard and behaviorally general ratings were compared (experimental condition 1), a clear-cut superiority was not apparent for either method. The mixed standard scale obtained slightly more significant correlations and the number of higher magnitude correlations was evenly split between the two rating methods. In summary, for first line ratings the behaviorally specific scales correlated better with the selected test scores than did either of the other two methods. There was little or no difference between the results obtained for the mixed standard and behaviorally general formats.

The results were similar for the second line ratings (Table 29). The behaviorally specific scale method showed a slight edge over the behaviorally general scale and a greater improvement in comparison to the mixed standard scale method

(again in terms of number of higher magnitude correlations, but not in the number of significant correlations).

Taking the results for first and second line raters as a whole, the data show that the behaviorally specific scales tended to correlate higher with the predictor scores and more often reached a level of statistical significance than did either the behaviorally general or mixed standard scale ratings. The superiority of the behaviorally specific scale was greater in comparison to the mixed standard scale than with the behaviorally general scale. The results from comparison of the behaviorally general and mixed standard scales showed little or no difference in the two formats.

CHAPTER V

SUMMARY AND DISCUSSION

The objective of this study was to contrast three rating methods, two of which have been recipients in the research literature of similar plaudits from their respective supporters. The experimental design was such that the effects of using a disguised or obvious scale continuum could be assessed along with the utility of behaviorally specific versus behaviorally general scale anchors. Differences between methods were apparent for various psychometric characteristics of the scales. To declare one rating method as superior to another, however, the full range of results obtained in that experimental condition must be considered.

Looking first at experimental condition 1, the comparison of mixed standard and behaviorally general scales, the results were consistent. No significant differences were found in the means of the ratings for the two methods nor in the correlations of the scales with test scores. Conclusions were not drawn concerning the standard deviations. The findings for the remainder of the scale characteristics, however, were in favor of the behaviorally general scale. These included: interrater reliability, all sources of variance as assessed by the multitrait-multirater analysis of variance, and factor pattern clarity obtained by factor analyzing the ratings. On the basis of these combined results, it would appear that the behaviorally general scale was superior to the mixed standard

scale method. Since these two rating methods shared identical behavioral anchors, the superiority of the behaviorally general scale is attributed to the obvious seven-point continuum on which these anchors were presented.

In experimental condition 2 the behaviorally general and behaviorally specific anchors were compared. Overall, the differences between these two methods were extremely small. No clear differences were evident in the means and standard deviations of the ratings. Results of the multitrait-multirater analysis, including interrater reliability, were also about equal. The behaviorally specific scales did, however, correlate better with the predictors. In addition, because of the manner in which the coefficients of congruence paired and the magnitude of these correlations, the behaviorally specific scale was judged to factor slightly better than the behaviorally general scale. On the basis of these last findings, overall results for the behaviorally specific method are seen as slightly more favorable than those for the behaviorally general method.

Had the mixed standard scale proven superior to the behaviorally general scale, then the comparison of the two prominent rating methods--the mixed standard scale and behaviorally specific scale formats--would have taken on additional importance. The results of this comparison in the third experimental condition, however, is only a repeat of the previous comparison of an obvious and a hidden scale continuum. No differences were found in the means of the two sets of

ratings and conclusions concerning the standard deviations were not drawn. The remainder of the results all favor the behaviorally specific method including: interrater reliability, multitrait-multirater analysis of variance, factor analysis, and correlation of the ratings with predictor scores. These findings indicate that the behaviorally specific method was consistently superior to the mixed standard scale method.

The overall results for the three experimental conditions leave little doubt that an obvious scale format is somewhat superior to the hidden continuum of the mixed standard scale. This study, therefore, supports Arvey and Yoyle's (1974) findings that disguising the continuity of the scale on which behavioral expectations are presented does not improve the ratings. Conversely, Blanz's (1965) contention that concealment of the final result of the evaluation from the rater will better reduce halo and leniency effects was not upheld.

While the advantages of highly specific behavioral expectations to more behaviorally general anchors are not as clear, the findings of the present study seem to fit into the large pattern of research already available on the superiority of the scaled expectations method (Campbell, Dunnette, Arvey, & Hellervik, 1973). The behaviorally specific scale was seen as slightly preferable based upon the results obtained in comparison to the behaviorally general scale. Supporters of the retranslation technique could effectively argue that the differences between the behaviorally specific scale and the other two methods would have been even more in favor of the specific

scale, had only raters using this scale been given the advantage of involvement in the retranslation procedure. Hakel (1971) found that training is a critical factor in the success of behaviorally based scales. Education and favorable attitudes developed from participation by raters in developing the scales may be one of the most important factors in the retranslation procedure. In the present study, any training of the raters as a result of this technique benefited all three rating methods.

The quality of the ratings obtained by all three methods was satisfactory, especially for the first line raters. With only one exception, these ratings obtained three meaningful factors, an appropriate number according to several factor-analysis studies (Ewart, Seashore, & Tiffin, 1941; Grant, 1955; and Rush, 1953). Furthermore, the ANOVA convergent and discriminant validity effects for all three methods were highly significant (although significant halo effects were also present in all three rating methods). Only the low interrater reliabilities obtained with the mixed standard scale method detracted from the quality of any rating method. Even this flaw was not a significant liability, as evidenced by the factor analysis results and predictor score correlations for the mixed standard scale method. Possibly raters in the mixed standard scale procedure made judgments upon different yet valid criteria. If this were the case, parallel ratings could possess low reliabilities while adequately sampling performance. Buckner (1959) conducted a study in

which test and training scores were more predictive of the less reliable ratings than of those with high agreement between raters. One explanation for this outcome and also for the results in the present study is that the ratings, while discrepant, represent a wider proportion of total job performance.

Since the results were reasonably favorable for all three rating methods, none can be excluded from future research. The practitioner, depending upon the purpose for which he conducts performance evaluations, may place priorities on different aspects of a scale. Since greater specificity of the behavioral anchors and use of the retranslation technique require expenditure of more time (and therefore greater cost) in development of the scales, the practitioner must weigh the advantages or disadvantages of each rating method separately for every situation.

One design limitation in this study which was previously touched upon briefly, is that it was not possible to counterbalance the presentation of the three rating methods and thereby control for order effects. Two rival hypotheses for the results obtained in this study are therefore plausible: (1) the quality of the first set of ratings in each experimental condition was enhanced by the novelty of the situation; i.e., the raters became bored and less conscientious on the second rating occasion, and (2) the quality of each second set of ratings was enhanced by any learning which occurred on the first rating occasion. Examination of the data tends

to indicate that, with the exception of the standard deviation of the ratings, there is little evidence that the position of the ratings affected the results. There were multiple instances where both first and second occasion ratings were superior on different rating characteristics. Furthermore, for the one rating method which was used on both occasions--the behaviorally general scale--the results for the two occasions were reasonably similar. The likelihood that a significant order effect bias existed was seen as unlikely.

Having shown that an obvious scale continuum is somewhat better than the mixed standard scale format, this author is not yet ready to abandon research with the mixed standard scale. The results discussed previously indicate that this scale does have merit as a measurement device. Furthermore, the possible advantage of the mixed standard scale in identifying rating errors argues for further study of this aspect of the method. Other studies have shown that characteristics of the raters: their personality (Guilford, 1954), their intelligence (Stockford & Bissel, 1949), and their own work effectiveness (Kirchner and Reisberg, 1962) all affect the evaluation of their subordinates. These are just a few of the variables which could be examined in relation to response errors on the mixed standard scale.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anastasi, A. Fields of applied psychology. New York: McGraw-Hill, 1964.

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

Barrett, R. S., Taylor, E. K., Parker, J. W. & Martens, L. Rating scale content: I. Scale information and supervisory ratings. Personnel Psychology, 1958, 11, 333-346.

Bass, B. M. Further evidence on the dynamic character of criteria. Personnel Psychology, 1962, 15, 93-97.

Bayroff, A. G., Haggerty, H. & Rundquist, E. A. Validity ratings as related to rating techniques and conditions. Personnel Psychology, 1954, 7, 93-113.

Bellows, R. M. Procedures for evaluating vocational criteria. Journal of Applied Psychology, 1941, 25, 499-513.

Bendig, A. W. Reliability and the number of rating scale categories. Journal of Applied Psychology, 1954, 38, 38-40. (a)

Bendig, A. W. Rater reliability and "Judgmental demoralization." Journal of Applied Psychology, 1957, 41, 66-68. (b)

Berkshire, J. & Highland, R. Forced choice performance rating: A methodological Study. Personnel Psychology, 1953, 6, 355-378.

Bingham, W. V., & Davis, W. F. Intelligence test scores and business success. Journal of Applied Psychology, 1924, 8, 1-22.

Blanz, F. Mixed standard scale: A new merit rating method. Its development and use in industry. Doctoral dissertation, Finland's Institute of Technology, Helsinki, 1965.

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

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.

Bray, D. W. & Moses, J. L. Personnel selection. Annual Review of Psychology, 1972, 23, 545-576.

Buckner, D. N. The predictability of ratings as a function of interrater agreement. Journal of Applied Psychology, 1959, 43, 60-64.

Buel, W. D. The validity of behavioral scale items for the assessment of individual creativity. Journal of Applied Psychology, 1960, 44, 407-412.

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

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

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

Campbell, J. P., Dunnette, M. D., Lawler, E. E. & Weick, K. E. Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

Campion, J. E., Greener, J. & Wernli, S. Work observation versus recall in developing behavioral examples for rating scales. Journal of Applied Psychology, 1973, 58, 286-288.

Cliff, N. Adverbs as multipliers. Psychological Review, 1959, 66, 27-44.

Comrey, A. L., High W. S. & Wilson, R. C. Factors influencing organizational effectiveness VII. A survey of aircraft supervisors. Personnel Psychology, 1955, 8, 245-257.

Cuomo, S. Validity information exchange, No. 8-17. Personnel Psychology, 1955, 8, 268.

Dunnette, M. D. A note on the criterion. Journal of Applied Psychology, 1963, 47, 251-254.

Edwards, A. Techniques of attitude scale construction. New York: Appleton-Century-Crofts, 1957.

Ewart, A., Seashore, S. & Tiffin, J. A factor analysis of an industrial merit rating scale. Journal of Applied Psychology, 1941, 25, 481-486.

Ferguson, L. The value of acquaintance rankings in criterion research. Personnel Psychology, 1949, 2, 93-103.

Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Flanagan, J. C. & Burns, R. K. The employee performance record: A new appraisal and development tool. Harvard Business Review, 1955, 33, 95-102.

Fleishman, E. A. & Kruchter, B. Factor structure and predictability of successive stages of learning Morse Code. Journal of Applied Psychology, 1960, 44, 97-101.

Folgi, L., Hulin, C. L. & Blood, M. R. Development of first-level behavioral job criteria. Psychological Bulletin, 1971, 55, 3-8.

Ghiselli, E. E. Dimensional problems of criteria. Journal of Applied Psychology, 1956, 40, 1-4.

Ghiselli, E. E. & Brown, C. W. Personnel and Industrial Psychology, New York: McGraw-Hill, 1955.

Ghiselli, E. E. & Haire, M. The validation of selection tests in light of the dynamic character of criteria. Personnel Psychology, 1960, 13, 225-231.

Gifford, W. W. Does business want scholars? Harper's Magazine, 1928, 156, 669-674.

Grant, D. L. A factor analysis of manager ratings. Journal of Applied Psychology, 1955, 39, 283-286.

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

Guion, R. M. Criterion measurement and personnel judgments. Personnel Psychology, 1961, 14, 141-149.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Guttman, L. A. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Hakel, M. D. Similarity of post-interview trait rating intercorrelations as a contributor to interrater agreement in a structured employment interview. Journal of Applied Psychology, 1971, 55, 443-448.

Harmon, H. H. Modern Factor Analysis. Chicago: University of Chicago Press, 1970.

Hartshorne, H., & May, M. A. Studies in service and self-control. New York: Macmillan, 1929.

Hausman, H. J. & Strupp, H. H. Non-technical factors in supervisors' ratings of job performance. Personnel Psychology, 1955, 8, 201-217.

Hays, W. L. Statistics for the social sciences. New York: Holt, Reinhart & Winston, 1973.

Henry, W. E. Executive personality and job success. American Management Association Personnel Series, 1948, No. 120.

Hulin, C. L. The measurement of executive success. Journal of Applied Psychology, 1962, 5, 303-306.

Jay, R. & Copes, J. Seniority and criterion measures of job proficiency. Journal of Applied Psychology, 1957, 41, 58-60.

Johnson, D. M. Re-analysis of experimental halo effects. Journal of Applied Psychology, 1963, 47, 46-47.

Johnson, D. M. & Vidulich, R. N. Experimental manipulations of the halo effect. Journal of Applied Psychology, 1956, 40, 130-134.

Jurgensen, C. E. Intercorrelations in merit rating traits. Journal of Applied Psychology, 1950a, 34, 240-243.

Jurgensen, C. E. Overall job success as a basis for employee ratings. Journal of Applied Psychology, 1950b, 34, 333-337.

Kavanaugh, M. J., MacKinney, A. C. & Wolins, L. Issues in managerial performance: multitrait-multimethod analyses of ratings. Psychological Bulletin, 1971, 75, 34-49.

Kay, B. The use of critical incidents in a forced-choice scale. Journal of Applied Psychology, 1959, 43, 269-270.

Kirchner, W. K. & Dunnette, M. D. Identifying the critical factors in successful salesmanship. Personnel, 1957, 34(2), 54-59.

Kirchner, W. K. & Reisberg, D. Differences between better and less-effective supervisors in appraisal of subordinates. Personnel Psychology, 1962, 15, 295-302.

Klieger, W. & Mosel, J. The effect of opportunity to observe and rater status on the reliability of performance ratings. Personnel Psychology, 1953, 6, 57-63.

Klores, M. A. Rater bias in forced-distribution performance ratings. Personnel Psychology, 1966, 19, 411-421.

Korman, A. K. Industrial and Organizational Psychology. Englewood Cliffs: Prentice-Hall, 1971.

Landy, F. J. & Guion, R. M. Development of scales for the measurement of work motivation. Organizational Behavior and Human Performance, 1970, 5, 93-103.

Lawler, E. E. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 369-381.

Lawshe, C. H., Kephart, N. C. & McCormick, E. J. The paired comparison technique for rating performance of industrial employees. Journal of Applied Psychology, 1949, 33, 67-77.

Leppowski, J. R. Development of a forced-choice scale for engineer evaluation. Journal of Applied Psychology, 1963, 47, 87-88.

Maas, J. B. Patterned scaled-expectation interview: Reliability studies on a new technique. Journal of Applied Psychology, 1965, 49, 431-433.

MacKinney, A. C. & Wolins, L. Validity information exchange. Personnel Psychology, 1960, 13(1), 443-447.

Mattell, M. A. & Jacoby, J. Is there an optimal number of alternatives for Likert-scale items? Journal of Applied Psychology, 1972, 56, 506-509.

McCormick, E. J. & Bachus, J. A. Paired comparison ratings. The effect on ratings of reductions in the number of pairs. The reliability of ratings based on partial pairings. Journal of Applied Psychology, 1952, 36, 123-127, 180-192.

Murray, H. A. Explorations in Personality. New York: Oxford University Press, 1949.

Oksala, O. A new method of merit rating for the facilitation of job evaluation. Proceedings of the XIII Congress of the International Association of Applied Psychology, 1958.

Ostle, B. Statistics in Research. Ames, Iowa: Iowa State University Press, 1964.

Otis, J. L. The criterion. In W. H. Stead, et al., Occupational Counseling Techniques. New York: American Book, 1940.

Peters, D. L. & McCormick, E. J. Comparative reliability of numerically anchored versus job-task anchored rating scales. Journal of Applied Psychology, 1966, 50, 92-96.

Richardson, M. W. & Kuder, G. F. Making a rating scale that measures. Personnel Journal, 1933, 12, 36-40.

Roach, E. D. & Wherry, R. J. Performance dimensions of multiline insurance agents. Personnel Psychology, 1970, 23, 239-250.

Robbins, J. E. & King, D. D. Validity information exchange. Personnel Psychology, 1961, 14, 217-219.

Ronan, W. W. & Prien, E. P. Towards a criterion theory: A review and analysis of research and opinion. Greensboro, North Carolina: The Richardson Foundation, 1966.

Rosensteel, R. K. A validation of a test battery and biographical data for machine operator trainees. Unpublished master's thesis, Bowling Green State University, 1953.

Ross, P. F. Reference groups in man-to-man job performance rating. Personnel Psychology, 1966, 19, 115-142.

Rotter, G. A. & Tinkleman, V. Anchor effects in the development of behavior rating scales. Educational and Psychological Measurement, 1970, 30, 311-318.

Rush, C. H. A factorial study of sales criteria. Personnel Psychology, 1953, 6, 9-24.

Sharon, A. T. & Bartlett, C. J. Effect of instructional conditions in producing leniency of two types of rating scales. Personnel Psychology, 1969, 22, 251-263.

Sisson, D. E. Forced-choice -- the new Army rating. Personnel Psychology, 1948, 1, 365-381.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In The Handbook of Industrial-Organizational Psychology, M. D. Dunnette, (Ed.) in press.

Smith, P. & 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.

Starch, D. An analysis of the careers of 150 executives. Psychological Bulletin, 1942, 39, 435.

Stockford, L. & Bissell, H. W. Establishing a graphic-rating scale. In Fleishman, W. E. (Ed.) Studies in Personnel and Industrial Psychology. Homewood, Ill.: Dorsey, 1967.

Stockford, L. & Bissell, H. W. Factors involved in establishing a merit rating scale. Personnel, 1949, 94-118.

Stoltz, R. E. Development of a criterion of research productivity. Journal of Applied Psychology, 1958, 42, 308-310.

Stoltz, R. E. Factors in supervisors' perceptions of physical science research personnel. Journal of Applied Psychology, 1959, 43, 256-258.

Taylor, E., Schneider, D. & Clay, H. Short forced-choice ratings work. Personnel Psychology, 1954, 7, 245-252.

Taylor, E. & Wherry, R. A study of leniency in two rating systems. Personnel Psychology, 1951, 4, 39-47.

Tiffin, J. Six merit rating systems. Personnel Journal, 1959, 37, 288-291.

Thurstone, L. L. Attitudes can be measured. American Journal of Sociology, 1928, 33, 529-544.

Thorndike, R. L. Personnel Selection. New York: Wiley, 1949.

Toops, H. A. The criterion. Educational Psychological Measurement, 1944, 4, 271-297.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Turner, W. W. Dimensions of foreman performance: A factor analysis of criterion measures. Journal of Applied Psychology, 1960, 44, 216-223.

Uhrbrock, R. S. & Richardson, M. W. Item analysis. Personnel Journal, 1933, 12, 141-154.

Vroom, V. H. Work and Motivation. New York: Wiley, 1964.

Wallace, A. R. Criteria for what? American Psychologist, 1965, 20, 411-417.

Ward, W. H. The "It's your business" approach to ratings. Personnel Psychology, 1961, 14, 183-190.

Wells, W. D. & Smith, G. Four semantic rating scales compared. Journal of Applied Psychology, 1960, 44, 393-397.

Williams, F. J. & Harrell, T. W. Predicting success in business. Journal of Applied Psychology, 1964, 48, 164-167.

Willingham, W. W. Interdependence of successive absolute judgments. Journal of Applied Psychology, 1958, 42, 416-418.

Yuzak, R. P. The Assessment of Employee Morale. Columbus: The Ohio State University Bureau of Business Research, 1961, Monograph No. 99.

Zedeck, S. & Baker, H. T. Evaluation of Behavioral Expectation Scales. Paper presented at the meeting of the Midwestern Psychological Association. Detroit, May, 1971.

Zedeck, S., Imparato, N., Krausz, M. & Oleno, T. Development of Behaviorally Anchored Rating Scales as a Function of Organizational Level. Journal of Applied Psychology, 1974, Vol. 59, No. 2.

APPENDIX A

Rating Instructions

RATING INSTRUCTIONS

1. This packet contains one form of the performance evaluations which you will use in assessing your Store Managers.
2. Before you begin the process of evaluating your people, you may find it helpful to review the following important points concerning ratings in general and this study in particular.

   If the purpose of the study were to help you decide which managers should be given pay raises or promotions, and which should be terminated, etc. you would want to consider which managers are your best and which are your poorest.

   However, the purpose of this study is not to evaluate managers for any of these purposes. Instead, this study is being done to identify which tests are best and which are poorest in selecting Store Managers. We therefore need to look at both the strengths and the weaknesses of every manager. Probably even your very best manager does some things wrong and your very poorest does many things well.

   With this in mind while making your ratings, you must be careful to:

   a) Temporarily disregard your overall impression of each manager; and

   b) Concentrate only upon the manager's performance on THE SPECIFIC PART of the job you are evaluating (This will help you avoid the error of

rating your best manager high on all parts and your poorest low on all parts of her job). By comparing both the strengths and the weaknesses of each manager with her test scores, we will be better able to identify these strengths and weaknesses in potential managers before they are hired.

3. Suggestions for completing the ratings:
   a) Try to work on the evaluation when you are most likely to be free of distractions and interruptions;
   b) Work on the evaluations for several short time periods rather than trying to do all the evaluations at once. When you become fatigued or bored, it is better to stop and continue later than to make rating errors;
   c) Evaluate every manager on one work part and then go on to the next work part (rather than evaluate one manager at a time on every work part);
   d) Please give each evaluation your utmost consideration, since the usefulness of the tests in helping you hire good managers depends upon your evaluating each person accurately.
4. Instructions for completing the rating forms:
   a) On the following pages you will find:
      (1) three pages each containing the names of all the managers you will be evaluating;

(2) a list of 57 descriptions of a manager's performance; and

(3) five rating forms each containing a 1 to 7 scale. Please check to make sure you have received the three pages containing the names of your managers, the list of 57 performance descriptions, and the five rating scales.

b) Beginning with performance description No. 1, read the description and then evaluate your first manager using the following system:

+ = manager is better than the description
0 = description fits the manager perfectly
- = manager is not as good as the description

Place a "+", "0", or "-" in the first square opposite the manager's name. Below is an example:

1. This manager is highly effective in managing her store. It is very difficult for another manager to equal her performance.

STORE MANAGER'S NAME	DESCRIPTION NUMBER												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Jane Smith	0												
Mary Jones	+												
Sarah Green	-												

Jane Smith is seen as fitting the description perfectly, so a "0" is placed beside her name for description No. 1. On the other hand, Mary Jones' performance is really better than the descriptive phrase so she is given a "+" on description No. 1. Now, evaluate your other store managers in similar fashion, using the following forms, until you have evaluated all your managers on the first description on page 7.

Next, read the second performance description on page 7 and repeat the above process, recording each manager's rating in Column 2. Continue until you have evaluated every manager on all 57 performance descriptions.