DOCUMENT RESUME

ED 237 913 CG 017 170

AUTHOR Tsui, Anne S.
TITLE Qualities of Judgmental Ratings by Four Rater Sources.
PUB DATE Sep 83
NOTE 52p.; A version of this paper was presented at the Annual Convention of the American Psychological Association (91st, Anaheim, CA, August 26-30, 1983). Support for this research was provided by the Duke Univeristy, Fuqua School of Business.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Administrator Evaluation; *Administrators; Error of Measurement; *Evaluation Methods; *Interrater Reliability; Measurement Techniques; Predictive Validity; Psychometrics
IDENTIFIERS Halo Effect; Leniency Response Bias; *Performance Appraisal; Restriction of Range

ABSTRACT

Quality of performance data yielded by subjective judgment is of major concern to researchers in performance appraisal. However, some confusion exists in the analysis of quality on ratings obtained from different rating scale formats and from different raters. To clarify this confusion, a study was conducted to assess the quality of judgmental ratings provided by four rater sources. Six indices which seem to be meaningful for assessing the quality of judgmental ratings by different raters, i.e., leniency, range restriction, halo, dimensionality, inter-rater agreement, and predictive validity, were used. Middle level managers (N=344) were judged on their managerial role effectiveness by 272 superiors, 606 subordinates, and 470 peers, who rated ten specific roles, three overall performance variables, and completed the company's formal performance rating. Results indicated that self-ratings were slightly more lenient, but had the least halo. Superiors' ratings had the most restricted ranges and the highest level of halo. Peer ratings contained less halo than the ratings by superiors and subordinates, had less restricted ranges than superiors' ratings, and showed some level of predictive validity. Subordinate ratings had the least restricted ranges, but more halo than self-ratings, and had the lowest predictive validity. There was low inter-rater agreement on the effectiveness ratings across all the rater sources. The results indicate the need for further research to provide a better understanding of the nature of ratings provided by different segments of an organization. (JAC)

Qualities of Judgemental Ratings by Four Rater Sources[1]

Anne S. Tsui

Fuqua School of Business
Duke University
Durham, N. C. 27706
(919) 684-3394

[1] Support for this research was provided by the Business Associates Fund of the Fuqua School of Business. A version of this paper was presented at the 91st Annual convention of the American Psychological Association, Anaheim, California, 1983.

September 1983

Qualities of Judgemental Ratings by

Four Rater Sources

ABSTRACT

There has been a long history of interest in the psychometric quality of subjective perfomance ratings. Some confusion, however, was found in the analysis of quality on ratings obtained from different rating scale formats and from different raters. This paper attempts to clarify this confusion by introducing six indices which seem to be meaningful for assessing the quality of judgemental ratings by different raters. The indices are leniency, range restriction, halo, dimensionality, inter-rater agreement and predictive validity. Using these six indices, the quality of judgemental ratings provided by four rater sources were evaluated and compared. 344 middle level managers were judged on their managerial role effectiveness by 272 superiors, 606 subordinates and 470 peers. Results indicate that self ratings are slightly more lenient, but have the least halo. Superior ratings have the most restricted ranges and the highest level of halo. Peer ratings contain less halo than the ratings by superiors and subordinates. They have less restricted ranges than superior ratings and show some level of predictive validity. Subordinate ratings have the least restricted ranges, but more halo than self ratings and have the lowest predictive validity. There is low inter-rater agreement across all the rater sources as well as between raters of the same source. Implications and suggestions for future theoretical and empirical work are discussed.

Qualities of Judgemental Ratings by Four Rater Sources

Introduction

Quality of performance data yielded by subjective judgement is of major concern to researchers in the performance appraisal arena, as evidenced by a number of articles addressing this issue in the recent years (e.g. Cooper, 1981; Kane & Lawler, 1979; Landy & Farr, 1979; Saal, Downey & Lahey, 1980). Saal, et al (1980) traced the historical development of theory and research involving the issue of measurement errors in judgemental performance appraisal data. They found researchers to have evaluated the quality of rating data by examining the degree to which they are free from halo, central tendency or range restriction, leniency or severity tendencies and show high inter-rater agreement or reliability. Other researchers have analyzed the dimensionality of the ratings (e.g. Landy, Vance, Barnes-Farrell & Steele, 1980), as well as their convergent and discriminant validities (e.g. Kavanagh, Wolins and McKinney, 1971). There is disagreement on whether inter-rater agreement is an appropriate measure of psychometric quality (Buckner, 1959; Lumsden, 1976). Borman (1974) suggested that the multitrait-multimethod analysis should not be used to assess the quality of ratings from different rater sources. It seems that there is some confusion in the literature regarding the appropriate indices for evaluating the quality of ratings by different raters. There is also a lack of knowledge regarding the quality of rating data from some rater sources, such as subordinates. It is the purpose of this paper to clarify some of the conceptual and operational confusion, to offer six indices as meaningful approaches for assessing the quality of rating data from multiple rater sources, and to apply these indices to a set of performance ratings given to a group of managers by their superiors, peers, subordinates and themselves. No research could be found in the literature that evaluates and compares the quality of performance ratings by all four rater sources.

Confusion Between Rating Quality and Construct Validity

The literature shows two parallel approaches to the study of quality in judgemental ratings. Beginning with Lawler (1967), the multitrait-multimethod procedure (Campbell-Fiske, 1959) was adopted to determine inter-rater agreement, discriminant validity and the extent of halo in the ratings (Heneman, 1974; Kavanagh, et al., 1971; Thomsom, 1970). Low inter-rater reliability, weak discriminant validity and a large halo effect are indicators of poor construct validity, i.e. low relevance to the ultimate criterion (Kavanagh, et al., 1971). Due to the lack of an ultimate criterion and the absence of a true performance score, another stream of research focuses its efforts on comparing the nature of rating errors in ratings data obtained from different rating scale formats and from different rater sources (see Saal, et. al., 1980 for a comprehensive review). Following Lawler's (1967) suggestion, subsequent researchers relied on the multitrait-multimethod approach (MTMM), treating both raters or rater sources and rating scale formats as alternative methods in obtaining performance ratings. This may have led to the confusion regarding the criteria for rating quality in terms of construct validity and in terms of freedom from error due to rater biases. Borman (1974) was one of the first researchers to point out the problem with using the MTMM procedure in assessing the ratings obtained from different raters. Convergent and discriminant validities are intended to assess the psychometric quality of the construct, not of the rater's responses. The quality of rating data obtained from different scale formats may be appropriately treated as different methods of measurement. Convergent validity among these ratings can be appropriately interpreted as indication of construct validity. The MTMM analysis of these ratings may also yield information on the extent of rating errors that are present in different rating scale formats. MTMM analysis, however, does not provide information on the relative accuracy, validity and reliability of the

ratings provided by different raters. It provides no knowledge on the relative degree of halo, leniency, or range restriction in the ratings among the raters (Borman, 1974). In fact, low convergent validity or inter-rater agreement may be meaningful when there are valid reasons for the divergent opinions among the multiple evaluators. This viewpoint has been expressed by many researchers (Borman, 1974; Buckner, 1959; Freebert, 1969; Tsui & McGregor, 1982; Wherry, 1952).

Evaluation of the psychometric quality of ratings from different rater sources, therefore, must rely on approaches other than the MTMM procedure. Unfortunately, neither the ANOVA procedure proposed by Kavanagh, et al., (1971) nor the MANOVA method suggested by Saal, et al., (1980) is sufficient to provide the information necessary for comparing the quality of ratings from different rater sources. The ANOVA procedure does not indicate which rater source, if there are more than two, have more leniency, or more halo. The MANOVA method as described by Saal, et al (1980) provides an overall leniency effect among the multiple sets of ratings. The halo effect is inferred from the number of latent roots. Saal, et al., (1980) further specifies that a significant ratee main effect may be interpreted as the absence of range restriction. However, this MANOVA approach is appropriate only when there is a full design in which all the raters rate all the ratees on all the performance dimensions. The ratee and the halo effects are confounded when only a partial design is used, in which blocks of raters rate some or perhaps only one of the ratees on all the performance dimensions. This is a condition not too dissimilar to performance evaluation of individuals in organizations. Under this condition, this MANOVA method is limited to only providing information on an overall leniency effect. Saal, et al., (1980) implied that a partial design in which ratees are nested under the raters may be used to perform the MANOVA. This application of the MANOVA model, however, is not

appropriate since it is a mixed model in which the rater is the fixed effect while the ratee is the random effect. The use of the ratee dimension as a main effect may not be appropriate in this partial design.[1] The MANOVA, however, is useful for providing an overall index of leniency by a significant rater main effect on the level of mean ratings.

Indices of Rating Quality from Multiple Raters

There is a growing body of literature advocating the view that there may be important and valid reasons for divergence in the ratings from multiple raters (Borman, 1974; Kane & Lawler, 1979; Klimoski & London, 1974; Latham & Wexley, 1979; Miner, 1968; Tsui, in press; Zammuto, London & Rowland, 1982). For managers, the potentially meaningful raters are self, superiors, peers and subordinates (Lawler, 1967; Tsui, in press). Multiple-rater studies of managerial performance are numerous (see Tsui & McGregor, 1982 for a review of these studies). These studies consistently found low convergence of ratings among the multiple raters. These studies also found that raters of the same level tend to agree more than raters from distant levels (Kavanagh, et al., 1971; Lawler, 1967; Thomson, 1970). Thus, there may be a need to differentiate the rater from the rater source as independent evaluators. The relationship of the rater to the ratee may suggest meaningful differences in performance expectations and in the criteria that are important to different raters (Borman, 1974; Tsui, in press). Thus, the level of inter-rater agreement may reflect differences in performance criteria used in the evaluation, differences in aspects of performance or behavior observed by the different raters, differences in performance information available to the evaluator, or unique rater differences in rating behavior. Low convergent validity may be attributed to be rating error only after the first three explanations are considered. The inter-rater agreement is a special case of intra-class correlation in which the question is the interchangeability of the

judges (Shrout & Fleiss, 1979). Presumably, judges from the same class or source would be more interchangeable than raters from different sources. Controversy on halo as an index of error came into the foreground with Cooper (1981). Until then, halo had always been treated as an invalid source of variance, leading to an unwanted degree of high intercorrelation among the ratings on multiple dimensions obtained from the same rater source or rating method. In the context of a multiple rater approach to evaluating performance, halo may be a meaningful index for comparing the extent to which each of the multiple sets of performance data is characterized by a generalized impression. The rater who is most intimately aware of the different performance areas should provide ratings that contain the least halo. Raters who are in the poorest position to observe the ratee's job behavior may tend to rely more on generalized impressions and their ratings on the multiple dimensions may be more highly intercorrelated with each other. Consequently, ratings from a rater source that contain more halo will be less useful in terms of performance feedback on the ratee's specific strengths and weaknesses. The dimensionality of the ratings on multiple performance factors will also differ among the multiple raters, depending in part on the extent to which judgement is affected by overall impressions. Ratings affected by generalized impressions of effectiveness will tend to improve in dimensionality when the variance of this overall impression is partialled out of the rating data. Dimensionality will not change if the ratings are not affected by overall impressions. Landy, et al. (1980) increased the number of factors in a set of ratings from three to six after partialling out the variance in an overall performance variable. The dimensionality of superior ratings, in particular, is improved when this effect is removed (Holzbach, 1978; Thomson, 1970). Thus, analysis of the dimensionality of rating data will yield information that may be important for understanding which rater

source relies on generalized impressions in their judgement of performance and which set of ratings is better differentiated in terms of its underlying structure.

Researchers are concerned with the leniency effect when it reflects biased perception by the rater. Self-ratings are found to be higher than ratings from superiors or peers by some researchers (Holzbach, 1978; Thornton, 1978; 1980) but not by others (Heneman, 1974; Thomson, 1970). A rating higher or lower than the individual warrants is a meaningful concern only when there is a true performance score for validating the rating given. In the absence of a true score, leniency is meaningful primarily as a relative index of rating tendencies among multiple raters. Tendency to rate oneself higher than other raters could be potentially problematic in terms of the ratee's acceptance of a lower rating by superiors or other raters. It is important to identify those specific raters who tend to be more lenient than others and those conditions which may generate or encourage more lenient ratings by some raters. For example, preliminary evidence by Thornton (1968) suggests that executives who tended to overrate themselves were found to be the ones who were considered least promotable on the basis of a criterion measure of success in the organization. Would there be a tendency to under-rate by those executives who were considered most promotable according to the organizational success criterion? Instead of evaluating an absolute standard of leniency as measured by deviation from the mid-point of the rating scale, leniency is more meaningful as an index of rating quality by comparing the mean rating levels across multiple raters.

The usefulness of performance ratings is also reduced when the raters do not discriminate among different ratees in terms of their respective performance levels. The ratings tend to cluster around a narrow section of the scale continuum, resulting in small variance or standard deviations.

Restriction of range on a measure is a psychometrically undesirable characteristic because it tends to depress its correlation with other measures. Superior ratings were found to be more restricted in range than self or peer ratings by Heneman (1974) and Thomson (1970). Both Holzbach (1978) and Thornton (1980), however, reported more range restriction in self ratings. Again, these inconsistent findings suggest the need for further investigation of the rating quality among multiple rater sources. Further conceptualization is also needed on the conditions under which one rater source may provide a more restricted range in ratings than other rater sources.

The quality of performance rating data may also be evaluated on the extent to which they relate to a common criterion on to which they meet the purposes intended. There are generally two purposes for obtaining performance ratings. First, the ratings are used to provide feedback for future effectiveness in job behavior; and second, they are used for determining rewards to be given to the ratee (Latham & Wexley, 1979). The psychometric quality of rating data should include an assessment of their predictive validity, i.e., the extent to which the ratings are predictive of rewards or of future performance. Kane and Lawler (1978) reviewed the peer assessment literature and concluded that peer ratings are most useful for providing feedback to the ratee for future performance improvement, while peer nominations are most predictive of the ratee's future promotions. Hegarty (1974) found feedback from subordinate ratings is associated with performance improvement of the supervisors ten weeks later. Researchers who advocate the multiple-rater approach to evaluating managerial performance rely on the argument that different raters would provide performance data that is unique, but of equal value to the person being evaluated. Empirical data is needed to investigate which ratings from multiple raters are more effective or valid in predicting a common performance criterion.

In sum, six different indices have been introduced and discussed to be meaningful for comparing the quality of judgemental ratings from multiple raters. These six indices are leniency, range restriction, halo, dimensionality, inter-rater agreement and predictive validity. While some of these indices may also be used to assess the construct validity of the performance measures, indices useful for the latter such as discriminant validity may not be equally meaningful for comparing the quality of rating data by multiple rater sources. An analysis of rating data across rater sources on these indices will provide useful information for evaluating the quality of the various sets of rating data. However, such information should not be directly interpreted to mean more or less accurate data from any specific source. Accuracy can be assessed only when a true performance score is available. The interest of this paper is not accuracy per se, but the differential qualities of judgement made on the performance of ratees by different raters.

Multiple Raters and Multiple Sources

Potential raters for an individual in the organization are many. Lawler (1967) suggests that superiors, peers, subordinates and self are all potentially meaningful raters for managers. Empirical studies of managerial performance have used raters such as assessment center staff members (Albrecht, Glaser & Marks, 1964), psychologists (Dicken & Black, 1965), training course faculty leaders (Mandell, 1957), subordinates (Hegerty, 1974; Kavanagh, et al; 1971), peers (Holzbach, 1978; Lawler, 1967), and second level superiors (Oldham, 1976). The majority of the raters, however, are the immediate superiors (see Lazer & Wikstrom, 1979 and Tsui, 1983 for reviews). The position of the rater relative to the ratee has been found to be one factor that needs to be considered in interpreting the convergent validity coefficients among multiple raters. Raters have a tendency to observe only

the behavior that relates to their own criteria for effective performance (Borman, 1974). A similar idea was proposed by Tsui (in press) who argues for a multiple-constituency approach to evaluating managerial effectiveness. The primary constituencies for a manager are the superiors, the subordinates and the peers. Thus, each constituency may be treated as a rater source. She suggests that criteria for managerial effectiveness tend to differ more across constituencies than between raters of the same constituency. This means that performance ratings should have a higher level of convergence between raters of the same source than between raters of different sources. Inter-rater agreement may then be used to assess the degree to which differences in ratings is due to different criteria held by different sources or due to unique rating behavior by different raters. Multiple rater studies may be as important and useful as multiple source studies.

The psychometric quality of superior and self ratings is more frequently studied than peer or subordinate ratings. Both peer and subordinate ratings have been shown to have predictive validity (Hegarty, 1974; Kane & Lawler, 1978; Roadman, 1964). Less is known about the degree of leniency, range restriction, halo or dimensionality in the ratings of these two rater sources, compared to the ratings of self and superiors. An empirical study was conducted to assess the quality of judgemental ratings provided by four rater sources, using the six indices described in this paper.

Method

Sample

The initial sample for this study consisted of a 10% stratified random sample of white male, and 50% of women and minority managers occupying positions ranging from second level section managers to vice presidents of a multi-divisional corporation. This sampling resulted in 550 managers. The purpose of the stratification was to ensure sufficient representation in the

four major functional areas of manufacturing, marketing, research and development, and administrative support services. The increased sampling percentage for women and minorities was to ensure a sufficient number to evaluate any potential race and sex differences in the analyses performed. It also permits the organization to perform special analyses pertaining to these two groups of managers. Of the 550 managers contacted, 344 responded to the study, resulting in a 62.5% response rate. In this sample, there were 217 white males, 78 white females and 49 minority managers.

Also participating in this study were 272 of these managers' superiors (79.1% response rate), 606 of their subordinates (88.1% response rate), and 470 of their peers (68.3%). The demographic characteristics of all the participants are summarized in Table 1. The four groups differ on all five demographic variables. The superiors are older, have the longest company and job tenures, have the highest educational level and are predominantly male. The subordinates are the youngest, have the shortest company tenure, the lowest educational level and have the largest proportion of females. The focal managers and the peers are similar on most of the demographic variables with the exception of job tenure and proportion of females. The largest proportion of females among the focal managers reflects the sampling ratio used. The lower job tenure also is related to the larger proportion of women since women in general were found to have lower company and job tenures.

Insert Table 1 about here

Procedure

After the initial sample was selected, a letter of introduction was sent to each manager by the corporation's vice president for personnel and public affairs. The research project was endorsed and the manager's cooperation was solicited. A week after this letter, a packet of six questionnaires was sent

to each manager by the researcher. The purpose of the research and the matter of confidentiality were explained in detail in the cover letter. Instructions for the disposition of the questionnaires were contained in both the cover letter and in the focal manager's questionnaire. The focal manager was asked to complete a white questionnaire and distribute the remaining five, which were printed in blue paper. Three criteria were given for the distribution of the blue questionnaires. First, the manager must give the questionnaires to one superior, two subordinates and two peers. They can be in either direct or indirect reporting relationships. Due to the matrix organizational structure, some of these managers worked more closely with people outside of their formal chain of command than those within it. Thus, a second criterion was that the manager must select those raters with whom he/she interacted most frequently on job-related matters. A third criterion was that the manager must select one subordinate and one peer with whom he/she worked best and one of each with whom he/she worked least well. The confidential nature of this research and the importance of these three criteria were strongly emphasized. A subsequent telephone call to each of 45 randomly selected managers indicated that they were comfortable with the research purpose and process, and that they all followed the criteria given for the selection of the raters. All raters independently provided the ratings in a confidential questionnaire which was mailed directly to the researcher in a self-addressed, stamped envelope provided.

Performance Measures

Performance was evaluated on ten managerial role behavior categories and one overall performance variable. The ten role behavioral categories were based on Mintzberg's managerial roles (1973). They are the representative, leader, liaison, environment monitor, information disseminator, spokesperson, entrepreneur, crisis handler, resource allocator, and negotiator roles. A

brief description of each role, which consists of a short paragraph with three to four sentences, was developed by Alexander (1979). For each role, the respondent was asked on a 7-point Likert-like scale the extent to which the manager was effective in performing the described role. The anchor ranged from (1) not at all, to (7) to an extreme extent. The wordings were appropriately modified for self-ratings.[2]

In addition to the ten specific role scales, each rater was also asked to respond to three questions comprising an overall performance variable, the expectational effectiveness scale used by Tsui (1982). It measures the extent to which the rater feels that the manager is performing his/her job consistent with the rater's expectations. The three items are (1) Overall, to what extent do you feel the focal manager is performing his/her job the way you would like it to be performed? (2) To what extent has he/she met your own expectations in his/her managerial roles and responsibilities? (3) If you entirely had you way, to what extent would you change the manner in which he/she is doing the job? Internal constitency reliability was estimated separately for each rater source. The alpha coefficient for the standardized items is α=.80 for self-ratings, α=.89 for ratings by the superiors, α=.88 for subordinates' ratings, and α=.87 for the ratings by peers. The average score of the three items was used for each rater source.

Finally, the company's formal performance appraisal rating was used as the criterion for estimating the predictive validity of the ratings on the ten managerial roles and on the expectational effectiveness scale. Two formal performance appraisal ratings were collected. The first was obtained at the same time all the raters were surveyed. The manager was asked to report the rating he/she received in the most recent performance appraisal given by the hierarchical superior. This appraisal was given within twelve months of

this research. A second formal performance appraisal rating was obtained eighteen months later. In a follow-up survey, the managers were asked to report the most recent formal appraisal rating that they had received from their hierarchical superior. These second ratings were obtained from a total of 257 managers who responded to the follow-up survey. The first formal rating was obtained from all 344 managers. This formal rating was measured on a 9-point scale, ranging from (1) below expectations, to (9) exceeds expectations. The rating is a summary score of the manager's effectiveness in work behavior as well as in accomplishing specific performance goals. The company uses this rating for administrative decision making regarding merit, promotions and transfers.

Conceptual and Operational Definitions

Based on the review of Saal, et al. (1980) and the conceptual meaning of the various indices, the following conceptual and operational definitions are used. Leniency is defined as the extent to which one rater source provides higher ratings on a set of performance dimensions than other rater sources. It is operationalized through the mean ratings on the eleven performance scales. Range restriction is defined as the extent to which a rater source has a tendency to give similar ratings to all ratees. It reflects a lack of discrimination on the performance levels among ratees. Range restriction is operationalized through the standard deviations on the performance scales of each rater source. Halo, on the other hand, is defined to be the extent to which a rater source is likely to give similar ratings to all the performance scales for a ratee. It reflects a lack of discrimination on the performance level across multiple traits. It is operationalized through the degree of inter-correlations among the performance variables within each source. Halo is more likely to occur when a rater is influenced by a generalized impression

about the ratee. Thus, dimensionality in the ratings can be improved when this generalized impression is controlled. Halo can therefore be further inferred by the change in the dimensional structure within a set of ratings by a rater source before and after the variance of an overall effectiveness measure is controlled for.

Dimensionality is defined as the extent to which the ratings are represented by a complex or simple underlying structure. It is operationalized through the factor pattern in the ratings given by a specific rater source. Differences in dimensionality among the rater sources are reflected in the different factor structures that may emerge in the ratings. Factor analysis of the residual variance after partialling out the overall effectiveness measure will reflect the extent to which a particular rater source may be more affected by an overall impression than other rater sources. Inter-rater agreement is defined as the extent to which two raters or two rater sources concur on the performance effectiveness of the same ratee. It is operationalized through the convergent validity coefficients of the monotrait-heteromethod diagonals. They are the pearson product-moment correlations of two raters on the same performance scale for the same ratee. Finally, predictive validity of the performance ratings is defined to be the extent to which the ratings of a rater source predicts the ratee's overall job success. It is operationalized through the multiple correlation between the performance scale ratings as predictors and the formal performance rating as the criterion measure.

Analysis

A preliminary two-way ANOVA was first performed on the eleven performance variables to test for potential race and sex effects. The race and sex of the managers (white male, white female and minority managers) were treated as a "sample" main factor. The rater source was treated as the "source" main

factor. A significant sample by source interaction effect would mean that the four sources have a tendency to evaluate the three groups of managers differently. The presence of this interaction effect would introduce error into the analysis of the pooled samples. The two-way ANOVA indicate no interaction effect on any one of the 11 performance variables. Thus, the three samples were pooled and the analyses of rating quality were performed on this total sample of 344 managers.

To measure the presence of leniency among the four rater sources, an overall MANOVA was first performed on the eleven performance scales. Then, a group by performance variable repeated measure MANOVA design was performed to evaluate the rater source by measure interaction effect. Eleven univariate F tests were also performed to assess the extent of leniency on each performance scale across the four rater sources.

Following a procedure recommended by other researchers, a non-parametric statistic was used to assess the probability that two sets of scores are given by subjects from two different populations. Heneman (1974) and Thomson (1970) used the sign test to establish that the two conditions are different. The sign test is appropriate in the case of two related samples. It analyzes the probability that one set of scores will be greater than another set by over 50% of the cases. This test has the most relaxed assumptions regarding the distribution of differences and the populations from which the different groups of subjects are drawn. The only requirement is that the variables being compared are continuous in measurement (see Siegel, 1959, pp. 68-72). The sign test, however, does not take into consideration the magnitude of the difference between the scores. The Wilcoxon matched-pairs, signed-ranks test has the same advantages of the sign test but takes into account the magnitude of the differences (Siegel, 1959, pp. 75-83). Thus, the Wilcoxon test was used to assess overall leniency by comparing the mean ratings on the eleven

performance scales by each rater source to the means of each of the other rater sources. The Wilcoxon test was also performed on the four sets of standard deviation scores to evaluate the relative degree of range restriction across the four rater sources. The level of halo in the four sets of ratings was assessed by applying the Wilcoxon test to the 55 inter-correlations on the eleven performance scales. Each set of 55 correlations was compared to each of the other three sets, resulting in six Wilcoxon tests. The expectational effectiveness measure was also included in the analysis to provide a more conservative test of halo since overall effectiveness is more likely to be affected by halo than ratings on specific performance variables.

Dimensionality of the ratings is estimated by factor analysis of the ratings on the ten managerial roles for each rater source. Mintzberg (1973) conceptualizes three general dimensions underlying these ten roles. Thus, factor analysis should yield the dimensions that he hypothesized. The expectational effectiveness variable was excluded from the factor analysis since it is not an integral part of the specific role dimensions for the managerial job. Principal component with varimax rotation was used. Kaiser's criterion was used in extracting the number of factors in the analysis of the ratings of each separate rater source. In these analyses, factors with eigen value of 1.0 or greater were retained and rotated. The effect of overall performance impression on the dimensionality of the rating data for each source was evaluated by the partialling procedure used by both Holzbach (1978) and Landy, et al., (1980). The expectational effectiveness score was used to partial out potential halo and the residual variance was then factor analyzed. Kaiser's criterion was also applied in determining the number of factors to retain and rotate. The coefficient of congruence or concordance was computed for the various sets of factor scores.

Inter-rater agreement was analyzed by pearson product-moment correlations on the ratings between two raters and between two rater sources on the same performance scale for the same ratee. Different procedures have been used to assess inter-rater agreement, including the average correlation, intra-class correlation and the Spearman-brown formula. Jones, Johnson, Butler and Main (1983) suggest that the average correlation is most informative if the researcher wants to know if different individuals are using the same decision rules in applying the rating scales. Finally, predictive validity of the ratings of the four rater sources was estimated by multiple regressions, using the eleven performance ratings as predictors and the two formal company performance ratings as the criterion measures.

Results

Leniency

The means and standard deviations of performance ratings by the four rater sources are summarized in Table 2. A significant overall MANOVA F was obtained ($F=12.31$, $p<.001$). Further analyses, however, show that the significant effect was not due to differences in the mean ratings across the four groups ($F=1.08$, N.S.) but due to differences in ratings over the eleven measures ($F=60,162.82$, $p<.0001$) as well as due to source-measure interactions ($F=10.58$, $p<.001$). Significant univariate F values were obtained on eight of the eleven scales. They indicate that self-ratings are highest on the leader, information disseminator, entrepreneur, crisis handler and resource allocator roles. These managers rated themselves lowest on the expectational effectiveness measure. Superiors rated these managers lowest on the spokesperson and entrepreneur roles. Subordinates gave the lowest ratings on the resource allocator role. Peers provided the lowest ratings on the leader, the information disseminator, and the crisis handler scales. These rating patterns indicate that leniency is observed on some of the performance

scales. A particular rater source may rate one performance scale higher but another lower. No one rater source provides consistently high ratings to all the performance scales, as would be suggested by an overall leniency effect. The lack of group effect in the MANOVA test was corroborated by the Wicoxon matched-pairs signed-ranks tests, performed on the means of the eleven variables across the four groups. In summary, these results show that leniency is a more complex phenomenon than suggested by past researchers. While there is an overall leniency effect, no one rater source gave consistently high ratings on all performance measures.

Insert Table 2 about here

Range Restriction

The four sets of standard deviations were compared for range restriction using the Wilcoxon test. The results show that the range of self-ratings are no more or less restricted than superior, subordinate or peer ratings. Superior ratings, however, have a greater range restriction than both the subordinate and the peer ratings ($T=7.00$, $p<.05$, and $T=8.00$, $p<.05$, respectively). As shown in Table 2, ten of the eleven standard deviations in the superior ratings are less than the correspondent standard deviations in both the subordinate and the peer ratings. Peer ratings have more range restriction than subordinate ratings ($T=7.50$, $p<.05$). Nine of the standard deviations in the subordinate ratings are larger than the standard deviations in the peer ratings. Thus, the results seem to suggest that the superior ratings have the most restricted ranges while the subordinate ratings have the least.

Halo

The heterotrait-monomethod correlations for each of the four rater sources are summarized in Table 3. The median correlation is the lowest for

the self ratings (r=.21). The Wilcoxon tests show that these correlations are much smaller than those in the superior (T=66, p.<001), the subordinate (T=39, p<.001), or the peer ratings (T=96.7, p<.001). The peer ratings have smaller correlations than both the superior (T=290.5, p<.01) and the subordinate ratings (T=260.0, p<.001). There is no difference in the magnitudes of the correlations between the superior and the subordinate ratings (T=706.5, n.s.). Thus, the results of this analysis indicate that self ratings may contain the least halo, followed by peer ratings. Superior and subordinate ratings seem to contain the largest amount of halo.

Insert Table 3 about here

A summary of the Wilcoxon tests for leniency, range restriction and halo in the ratings of the four rater sources is given in Table 4. In general, this set of results show that the overall leniency effect was due to high ratings on some performance scales by some rater source. Self ratings were found to have the highest score on five of the eleven performance scales and are higher than superior ratings on eight of the eleven variables. In fact, when the expectational effectiveness variables was not included in the Wilcoxon analysis, the difference between self and superior ratings became significant (T=7.5, p<.05). Thus, there is a tendency for self ratings to be more lenient than superior ratings. Superior ratings show the most restricted range, especially when compared to the ratings of subordinates and peers. Both superior and subordinate ratings had more halo than peer or self ratings. Self ratings appear to be least affected by halo. Subordinate ratings, however, have less range restriction than both superior and peer ratings.

Insert Table 4 about here

Dimensionality

The results of factor analysis on the performance ratings by the four rater sources are presented in Table 5. The factor structures and the loadings of both the raw scores and the residual scores are presented. Using Kaiser's criterion, two factors emerged in both the raw and the residual ratings for all four rater sources. However, the factor structures on self ratings is almost identical on both sets of ratings. The coefficients of congruence are above .980 on both factors. Two factors also emerged in the subordinate ratings on both the raw and the residual ratings. The coefficients of concordance for the two factors also exceed .980. Furthermore, the factors for the self and the subordinate ratings are highly congruent. The coefficients all exceed .960. The items defining each factor are identical for the self and the subordinate ratings. Thus, subordinates and the managers themselves seem to define the managerial job in a similar manner.

An examination of the items loadings on the two factors suggests that these managerial activities could be organized into internal and external roles. The first factor is defined by the representative, liaison, environment monitor, spokesperson, entrepreneur and the negotiator roles. These roles are externally-oriented. The second factor is defined by the leader, the information disseminator, the crisis handler and the resource allocator roles. These roles are internally-oriented. Thus, acccording to the managers and their subordinates, the managerial job as measured by these ten roles has an external and an internal dimension which do not replicate the three factors that Mintzberg postulated.

Different factor patterns were observed for the superior and the peer ratings. Though two factors emerged in both sets of the raw ratings,

according to Kaiser's criterion, only one factor is interpretable. All ten items load highly on one factor. The coefficients of congruence suggest that these two factor patterns are dissimilar to that of the self and the subordinate ratings (ranging from .779 to -.367). After controlling for overall effectiveness, however, two interpretable factors appeared in both the superior and the peer ratings. The two factors became more similar to that of the self and the subordinate ratings. The coefficients of congruence show that the two factors in the superior-ratings on the residual variance are more similar to those in the subordinate ratings (.965 and .938 for factor I and II, respectively) than those in the self-ratings (.830 and .922, respectively). The second factor in the peer-ratings is in concordance with the second factor in the self-ratings at .910. All the other cofficients of concordance exceed .940, a value which Tucker considers to be needed to define congruent factors (1951).

A closer examination of the factor loadings show that superiors see the entrepreneur role to be associated with both internal and external activities, by the high loadings on both factors. Also different from the perspectives of the other three rater sources, superiors see the negotiator role to be internal activities related to the management of subordinates while the other rater sources associate this role with external relations. According to the peers the information disseminator role is both an internal and an external activity. The remaining items have a loading pattern similar to subordinate and managers themselves. In general, the dimensionality seems to increase in complexity after overall impression of effectiveness is controlled for in the superior and peer ratings. They became more similar to the factor structures in the self and subordinate ratings. There was no change in the factor

structures of the self and subordinate ratings. Also, the data did not support the three general dimensions postulated by Mintzberg (1973). The ten roles appeared to be organized into external and internal activities in the cognitive maps of these rater sources.

Insert Table 5 about here

Inter-rater Agreement

Degree of inter-rater agreement was examined both across rater sources and between raters of the same source. With four rater sources, six sets of inter-source agreement correlation coefficients were computed. Two sets of intra-source inter-rater agreement correlations were computed, one for the subordinate and one for the peer source. These eight sets of agreement coefficients are summarized in Table 6. The correlations are low on both between and within rater source analyses. The highest correlations are between the superior and the peer ratings (median r=.23). This finding is consistent with earlier research (e.g. Lawler, 1967). The low level of agreement between the two subordinates and between the two peers is not surprising, given that the managers were instructed to select one subordinate and one peer with whom he/she worked best and one of each with whom he/she worked least well. In general, the managers received the lowest level of agreement in terms of their effectiveness on the environment monitor role from all the raters (median r=.065). Only one of the eight correlations reach significance. The highest level of agreement is found on the representative role (median r=.235) and on the expectational effectiveness variable (median r=.24). Overall, these correlation coefficients may be considered low. On the average, no two raters or two rater sources provide ratings that share more than 6% variance on the performance variable.

Insert Table 6 about here

Predictive Validity

Results of the regression analysis using the eleven performance scales ratings as predictors and the company's formal performance appraisal as the criterion measure are summarized in Table 7. Two regressions were performed for each rater source, one on the first formal performance rating and another on the second formal performance rating that the managers received from their hierarchical superiors. All four regressions for the first formal performance rating are significant, while only three of the four are significant on the second formal rating. The subordinate's regression model did not reach statistical significance for the second formal rating. The R^2s are the highest for the superiors regression models. The ratings on the eleven role performance scales account for 26% of the variance in the first formal company rating and 22% of the variance in the second formal rating. Predictive power is similar in strength in the self and the peer ratings. The eleven scores account between 7% and 10% of the variance in the formal ratings. The weakest predictive power is found in the subordinate ratings. Only 5% of the variance in the first dependent variable is accounted for by the ratings in the eleven role performance scales. Furthermore, the expectational effectiveness scale has significant beta weights in the regression models for superiors, peers and self, but not for subordinates. A low level of effectiveness in the environment monitor role and a high level of effectiveness in the entrepreneur role as evaluated by the focal managers themselves are predictive of formal rating given by the hierarchical superior. A low level of effectiveness in the environment monitor role as perceived by superiors is predictive of a high score on the second formal performance rating. None of the other roles

contribute to formal rating of effectiveness. Overall, the predictive validity is best for the superior ratings. This may be due to the fact that the formal performance rating was also given by most of these superiors, thus the high predictive validity may be a result of the common rater source, or common method variance. Some level of predictive validity in self-ratings is not surprising. A level of predictive validity in peer ratings is consistent with previous research findings (Barrett, 1964; Kane & Lawler, 1978; Kraut, 1975; Roadman, 1964).

Insert Table 7 about here

Discussion

Both researchers and managers are interested in the quality of performance rating data obtained from organizational members. Researchers are interested in their construct and predictive validities as well as their departure from the true score. Managers are interested in their usefulness for feedback and for rational administrative decision making. Thus, over the years, much research effort has been devoted to the analysis of alternative rating techniques for improving rating quality (e.g. Bernardin, 1977; DeCotiis, 1977; Keaveny & McGann, 1975; King, Hunter & Schmidt, 1980; Saal & Landy, 1977) and to evaluating the usefulness of training for reducing rating errors (e.g. Bernardin, 1978; Borman, 1979; Latham, Wexley & Pursell, 1975). Recently, attention has focussed on reviewing the meaningful indices of rating quality (e.g. Cooper, 1981; Saal, et al., 1980) and on assessing the quality of ratings by different raters or rater sources (e.g. Heneman, 1974; Holzbach, 1978). There has been some research in comparing the quality of superior, peer and self ratings, resulting in some inconsistent findings. However, there is no study to date that compares the ratings by all four rater sources for the same group of managers utilizing a common set of performance

variables. The research reported in this paper is intended to fill that gap in the literature.

Results show that the four rater sources differ in terms of their relative positions on the six indices. While overall, no one rater source was found to exhibit greater or less leniency across all eleven performance variables, there is a tendency for self ratings to be more lenient than superior ratings. This finding supports the results of research by Holzbach (1978) and Thornton (1968) but disagrees with the findings of Heneman (1974) and Thomson (1976). It is interesting to note that while the managers rated themselves higher than their superiors on their effectiveness in performing the ten managerial roles, they rated themselves lowest in meeting their self expectations of performance. Also, on some performance scales, higher ratings were given by the other raters. This rater by performance variable interaction suggests that the leniency effect is highly complex and that an overall leniency index is not sufficient to uncover the rating tendencies by different raters. Future research should explore the relative level of leniency on multiple performance dimensions. Further conceptualization is also needed to identify those individual or contextual variables that may lead to lenient or severe ratings by a particular rater or rater source. For example, Thorton's research (1968) suggests that there is a tendency for poor performers to give themselves more lenient ratings. Rater consequence may also affect the quality of the ratings (DeCotiis & Petit, 1977).

Superior ratings were found to have the most restricted ranges. This means that superiors tend to give similar ratings to their subordinate managers. It is therefore more difficult to discriminate the good performers from the poor ones based on the superior ratings. Subordinate ratings, on the other hand, seemed to have the least restricted ranges. This means that there are wider differences in the relative effectiveness or ineffectiveness among a

group of managers as perceived by their employees. The standard deviations in the subordinate ratings are larger than both those in the superior and peer ratings. On the other hand, peer ratings have better discrimination among variables than either the superior or the subordinate ratings. The highest degree of trait variance, however, was found in the self ratings. This result is meaningful since the managers themselves would be in the best position to have the most intimate knowledge on their relative strengths and weaknesses across a number of performance dimensions. Superiors appear to rely most on generalized impressions, resulting in the higher degree of inter-correlations among the various performance dimensions. The high degree of halo found in superior ratings is consistent with earlier research (Heneman, 1974; Holzbach, 1978).

Overall impression of effectiveness seems to affect the dimensionality of the superior and peer ratings. The dimensional structure is more complex after the variance of overall effectiveness is partialled out. While peer ratings contain less halo than either the superior or the subordinate ratings, the dimensional structure among the ten variables is further improved after the variance of overall impression is removed. It is interesting to note that the halo effect found in the subordinate ratings does not affect their dimensionality. A comparison of the four factor patterns suggests that superiors have a slightly different cognitive map of the ten managerial roles than the other three rater sources. The factors are most different from those in the focal managers' ratings. A role that was seen as external activities by the managers was viewed as internal by the superiors. This difference may have implications on the mutual understanding of role specifications by these two rater sources.

The partialling approach used in this study is not meant to suggest that halo is considered entirely as a measurement error. The author agrees with Cooper (1981) and Hulin (1982) that halo may contain an element of true score. The partialling procedure confirmed the finding by Landy et. al (1980) that overall impression affects the dimensionality of ratings by different sources. Also, overall impression of effectiveness may not be equivalent to halo in terms of its effect on the dimensionality of ratings.

The low degree of inter-rater agreement both across rater source and between raters of the same source is also consistent with previous research. The slightly higher degree of agreement between the superior and the peer ratings has also been found by previous researchers (e.g. Kavanagh, et al., 1971; Lawler, 1967). The weak agreement between the two subordinates and between the two peers, however, does not support the suggestion that raters from the same level would tend to evaluate more similarly than raters from different levels (e.g. Albrecht, et al., 1964; Gunderson & Nelson, 1966; Kavanagh, et al., 1971). The rater selection procedure used in this research might indicate the influence of contextual and/or interpersonal factors. The differential ratings given by multiple raters for the same manager may be explained by either an informational or a motivational perspective. Kane and Lawler (1979) offer primarily an informational explanation. Raters have different opportunities to observe the ratee's work behavior. What they see might not be a representative sample of the ratee's total spectrum of work activities. Borman (1974) and Tsui (in press), on the other hand, offer primarily a motivational explanation. They both suggest that raters use different criteria in evaluating the ratee's effectiveness. These criteria are based on the self interests of the raters. They are important to the rater's own work roles. Raters tend to observe other's behavior that relates

to their own criteria for effective performance. The differential ratings by different raters, whether they are within the same source or in different sources, may be a consequence of either or both informational and motivational factors. Those raters who may have similar opportunity to observe a ratee's work behavior and who may use similar criteria may evaluate similarly as well. The low correlations between the two raters within the subordinate or the peer sources suggests that factors other than expectations, such as interpersonal relationships, may also affect rating behavior. Heneman (1980) proposes that theory, measurement and behavioral focus should all be considered in explaining rating discrepancies among multiple raters or rater sources.

The predictive validity of the performance ratings by the multiple raters is modest in this research. Only two roles are predictive of the formal performance rating. They are the environment monitor and the entrepreueur roles. The environment monitor role describes the manager's behavior in processing written information and in keeping on top of the informal grapevine regarding decisions made by top management as well as trading gossip with industry contacts. Both the managers themselves and their superiors do not seem to value effective behavior on this role. They detract from getting a good formal performance rating. Two possible explanations can be offered for the overall weak relationship between effectiveness ratings on the ten roles and the formal rating. First, criteria other than behavior on the ten managerial roles may be important in determining the manager's worth to the corporation as assessed by the hierarchical superior. These may include criteria that relate to accomplishment of specific financial expectations or project achievements. Second, the formal performance rating itself may be contaminated with factors other than the manager's true performance. Indeed,

the criterion problem reviews is a thorny issue troubling researchers. Meeting the overall expectations of superiors, peers and self, however, contributes to a high formal ratings.

The shared variance between the peers' ratings on the ten roles and the formal rating seems to suggest that peer perceptions or opinions in this organization may be included in the superiors' judgement of the manager's total effectiveness or contribution. The regression results also show self-opinions affecting the formal rating, though to a modest degree only, as was in the peer ratings. The subordinates' opinions, however, did not seem to have any impact on the managers' formal review. Again, this finding may be explained by an informational perspective. Superiors may have more informational exchange with the focal manager's peers who may report to these superiors as well. The subordinates, however, are at least two hierarchical levels below. Subordinates having the least impact on the formal judgment of the manager's effectiveness may also reflect the formal authority structure. Subordinates are usually not involved in the formal performance appraisal of their supervisors. In fact, formal subordinate evaluation of superiors are almost non-existent in the literature. However, tentative research suggests that subordinate feedback may have positive effects on supervisory performance (Daw & Gage, 1967; Hegarty, 1974). There is a need to conceptualize and investigate those conditions under which subordinate ratings would have high predictive validity or informational value to the managers. It would be interesting to compare the validity of subordinate and peer assessment for predicting managerial promotion and success. Much more research is needed on subordinate feedback or ratings on their managers.

Conclusion

In summary, the results of this study suggest that different rater sources seem to possess different rating qualities. Self ratings seem to contain the least halo, but have a tendency to be more lenient than superior ratings. Superior ratings have the most restricted ranges and have the highest level of halo. Peer ratings have less halo than either the superior or the subordinate ratings but more restricted ranges than subordinate ratings. Both self and peer ratings have some level of predictive validity. Subordinate ratings have the largest variance in rating the group of managers. They, however, have high halo and low predictive validity. Self and subordinate ratings have a more complex cognitive structure of the ten managerial roles and the structures are not affected by overall effectiveness impression as are the superior or peer ratings. Finally, there is a low degree of agreement on the effectiveness ratings across all the raters.

Without a theoretical understanding of the causes for these differential qualities, it is inappropriate at this point to judge whether the quality of ratings from one rater source is better or worse than another. What this research suggests is that more theoretical and empirical work is needed for a better understanding on the nature of ratings provided by different segments of organizational members. Most research and almost all organizational practices still rely primarily on superior ratings when information about performance or when a performance criterion is needed (Lazer & Wikstrom, 1979). It seems that there are other potential raters who may provide performance ratings with qualities that may be more appropriate for certain research purposes or for managerial practice. Many researchers have advocated the use of multiple raters for both research and practice in the recent years (Kane & Lawler, 1979; Latham & Wexley, 1979). The multiple-rater approach

to assessment seems especially appropriate for managers (Miner, 1968; Tsui & McGregor, 1982). Before adoption of such practices, however, further research and conceptualization will be needed to explore the conditions under which one rater source may yield ratings with some desired qualities and the conditions under which one set of ratings from a particular rater source may lead to desirable organizational consequences.

Footnotes

[1]Personal communication with Eliott Cramer, original producer of the MANOVA computer program that Saal, Downey and Lahey (1980) recommend for use in analyzing the ratee main effect.

[2]This set of role descriptions was reported in Alexander's disertation. Since it is not easily accessible, it is encluded in the appendix of this paper.

Reference Notes

Tsui, A.S. The measurement of managerial effectiveness: progress and problems. Working paper, Fuqua School of Business, Duke University, 1983.

Tsui, A. S. & McGregor J. The multiple-rater approach to measuring managerial performance: further empirical evidence. Proceedings. The 14th annual meeting of the American Institute for Decision Sciences, 1982.

Wherry, R. L. The control of bias in ratings: VII. A theory of rating (Final Report No. 922). Department of the Army, Personnel Research Branch (now Army Research Institute for the Behavioral and Social Science) February, 1952.

REFERENCES

Albrecht, P., Glaser, E.M., & Marks, J. Validation of a multilple-assessment procedure for managerial personnel. Journal of Applied Psychology, 1964, 48(6), 351-360.

Alexander, L.D. The effect of level in the hierarchy and functional area on the extent to which Mintzberg's managerial roles are required by managerial jobs. Ph.D. Dissertation. University of California, Los Angeles, 1979.

Barrett R.S. Performance ratings. Chicago: Science Research Associates, 1966.

Bernardin, H.J. Behavioral expectation scales versus summated scales: a fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

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

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

Borman, W.C. The rating of individuals in organizations: an alternate approach. Organization Behavior and Human Performance, 1974, 12, 105-124.

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

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

Cooper, B. Ubiquitous halo. Psychological Bulletin, 1981, 90(2), 218-244.

Daw, R.W. & Gage, N.L. Effect of feedback from teachers to principals. Journal of Educational Psychology, 1967, 58(2), 181-188.

DeCotiis, T.S. An anlysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977, 19, 247-266.

DeCotiis, T. & Petit, A. The performance appraisal process: a model and some testable propositions. Academy of Mangement Review, 1977, 2(4), 635-646.

Dicken, C.F. & Black, J.D. Predictive validity of psychometric evaluations of supervisors. Journal of Applied Psychology, 1965, 49(1), 34-47.

Freeberg, N.E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Journal of Applied Psychology, 1969, 53, 518-524.

Gunderson, E.K.E. & Nelson, P.D. Criterion measures for extremely isolated groups. Personnal Psychology, 1966, 19, 67-82.

Hegarty, W.H. Using subordinate ratings to elicit behavioral changes in supervisors. Journal of Applied Psychology, 1974, 59(6), 764-766.

Heneman, H.G., III. Comparisions of self and superior ratings of managerial performance. Journal of Applied Psychology, 1974, 59(5), 638-642.

Heneman, H.G., III. Self-assessment: a critical analysis. Personnel Psychology, 1980, 33(2), 297-300.

Holzbach, R.L. Rater bias in performance ratings: superior, self and peer ratings. Journal of Applied Psychology, 1978, 63(5), 579-588.

Hulin, C.L. Some reflections on general performance dimensions and halo rating error. Journal of Applied Psychology, 1982, 67(2), 165-170.

Jones, A.P., Johnson, L.A., Butler, M.C. & Main, D.S. Apples and oranges: an empirical comparison of comonly used indices of inter-rater agreement. Academy of Management Journal, 1983, 26(3), 507-519.

Kane, J.S. & Lawler, E.E., III. Methods of peer assessment. Psychological Bulletin, 1978, 85, 555-586.

Kane, J.S. & Lawler, E.E. III. Performance appraisal effectiveness: its assessment and determinants. In Staw, B., Research in Organizational Behavior, 1979, 1, 425-478.

Kavanagh, M.J., MacKinney, A.C. & Wolins, L. Issues in managerial performance: multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 75(1), 34-49.

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

King, L.M., Hunter, J.E. & Schmidt, F.L. Halo in a multidimensional forced-choice performance evaluation scale. Journal of Applied Psychology, 1980, 65(5), 507-516.

Klimoski, R.J. & London, M. Role of the rater in performance appraisal. Journal of Applied Psychology, 1974, 59(4), 445-451.

Kraut, A.I. Predicting managerial success by peer and training ratings. Journal of Applied Psychology, 1975, 60(1), 14-19.

Landy, F.J. & Farr, J.L. Performance ratings. Psychological Bulletin, 1979, 87, 72-107.

Landy, F.J., Vance, R.J., Barnes-Farrell, J.L. & Steele, J.W. Statistical control of halo errors in performance ratings. Journal of Applied Psychology, 1980, 65, 501-506.

Latham, G.P. & Wexley, K.N. Increasing Productivity through Performance Appraisal. Reading, MA: Addison-Wesley, 1981.

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

Lawler, E.E., III. The multitriat-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 396-381.

Lazer, R.L. & Wikstrom, W.S. Appraising Managerial Performance: Current Practices and Future Directions. New York: The Conference Board Report, 1979.

Lumsden, J. Test theory, In Rosenzweiz, M.R. & Porter, L.W. (eds.) Annual Review of Psychology, (Vol. 27). Palo Alto, Calif: Annual Reviews, 1976.

Mandell, M. The selection of executives. In Dooher, M.J. & Marting, E. (eds.) The Selection of Management Personnel, I & II. New York: American Management Association, 1957.

Miner, J.B. Managerial appraisal: a capsule review and current references. Business Horizons, 1968, 11(5), 83-96.

Mintzberg, H. The Nature of Managerial Work. New York: Harper & Row, 1973.

Oldham, G.R. Motivational strategies used by supervisors: relationships to effectiveness indicators. Organizational Behavior and Human Performance, 1976, 15, 66-86.

Roadman, H.E. An industrial use of peer rating. Journal of Applied Psychology, 1964, 48, 211-214.

Saal, F.E., Downey, R.G., & Lahey, M.A. Rating the ratings: assessing the psychometric quality of rating data. Psychological Bulletin, 1980, 88(2), 413-428.

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

Shrout, Patrick E. and Fleiss, Joseph L. Intra-class Correlations: Uses in Assessing Rater Reliability. Psyschological Bulletin, 1979, 86(2), 420-428.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

Thomson, H.A. Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. Journal of Applied Psychology, 1970, 54(6), 496-502.

Thornton, G.C. The relationship between supervisor and self appraisals of executive performance. Personnel Psychology, 1968, 21(4), 441-455.

Thornton, G.C., III. Psychometric properties of self-appraisals of job performance. Personnel Psychology, 1980, 33(2), 263-271.

Tsui, A.S. A role set analysis of managerial reputation. Proceedings. The 42nd annual national meeting of the Academy of Management, 1982. Also to appear in Organizational Behavior and Human Performance.

Tsui, A. S. A Multiple-constituency framework of managerial reputational effectiveness. Presented at the NATO-sponsored International Symposium on Leadership and Management. To be published in Hunt, J.J., Hoskin, D., Schriesheim, C. & Stewart, R. (Eds). Leadership, Vol. 7, New York: Pergamon, in press.

Tucker, L.R. A method for synthesis of factor analysis studies. Personnel Research Section Report, No. 984, Washington, D.C.: Department of the Army, 1951.

Zammuto, R.F., London, M. & Rowland, K.M. Organization and rater differences in performance appraisal. Personnel Psychology, 1982, 35, 643-658.

able 1 Sample Demographics

emographic ariable	Rater Sources: Self (N=344) $\bar{X}$	Self SD	Superiors (N=272) $\bar{X}$	Superiors SD	Subordinates (N=606) $\bar{X}$	Subordinates SD	Peers (N=470) $\bar{X}$	Peers SD	Significa test
ge (years)	41.54	8.21	44.14	7.44	37.11	8.95	40.66	7.91	F=51.96**
ompany Tenure (years)	10.94	6.75	13.65	6.99	8.59	6.73	11.63	7.70	F=36.67**
ob Tenure (years)	2.75	2.90	3.63	4.12	3.04	3.33	3.24	3.87	F= 3.32*
ducation (years)	15.80	1.95	16.30	1.65	15.27	2.11	16.00	2.00	F=21.74**
ex (% male)	74.4%		96.3%		66.0%		85.3%		X^2=121.08**

* p<.05
** p<.01
** p<.001

Means and Standard Deviations of Performance Ratings by Four Raters Sources

ormance riable	Self (N=344) Mean	SD	Superiors (N=272) Mean	SD	Subordinates (N=606) Mean	SD	Peers (N=470) Mean	SD	ANOVA F
resentative	3.73	1.63	3.79	1.47	3.81	1.51	3.78	1.53	.191
der	5.53	.95	5.13	1.09	4.94	1.27	4.80	1.19	29.333*
ison	4.64	1.25	4.62	1.16	4.78	1.23	4.77	1.22	1.893
ironment Monitor	4.50	1.22	4.33	1.09	4.51	1.18	4.48	1.16	1.635
ormation Dessem.	5.26	1.05	4.85	1.05	4.98	1.20	4.82	1.15	10.959*
kesperson	4.51	1.46	4.39	1.35	4.79	1.22	4.62	1.26	7.109*
repreneur	4.34	1.44	4.00	1.26	4.31	1.35	4.07	1.30	6.276*
sis Handler	5.82	1.03	5.25	1.15	5.20	1.28	5.07	1.25	28.489*
ource Allocator	4.98	1.20	4.88	1.15	4.73	1.25	4.80	1.16	3.429*
otiator	4.48	1.55	4.65	1.29	4.54	1.42	4.59	1.32	.373
ectational ctiveness	4.71	1.03	5.03	1.07	5.07	1.23	5.02	1.13	8.024*
age over all ariables	4.77	.69	4.63	.77	4.70	.83	4.62	.78	2.70*

all MANOVA F = 12.31***
p Main Effect F = 1.08, n.s.
ures Main Effect F = 60,162.82***
p - measure interaction F = 10.58***

p <.05
p <.01
p <.001

Table 3 Heterotrait-Monomethod Correlations for Four Rater Sources

Self (N=344)

1	2	3	4	5	6	7	8	9	10	11
-										
22	-									
34	21	-								
29	20	36	-							
13	26	22	21	-						
42	15	43	31	13	-					
34	12	21	28	17	24	-				
20	24	18	13	24	20	17	-			
07	25	12	13	28	22	15	34	-		
35	18	36	20	10	39	27	19	30	-	
17	32	20	21	16	16	16	04	05	10	-

median r=.21

Superiors (N=272)

	1	2	3	4	5	6	7	8	9	10	1
1.	-										
2.	21	-									
3.	36	33	-								
4.	35	27	44	-							
5.	23	39	20	33	-						
6.	47	27	53	46	38	-					
7.	33	46	37	32	32	49	-				
8.	15	58	37	17	27	28	41	-			
9.	21	45	19	19	28	27	45	54	-		
10.	25	47	34	35	30	41	49	51	45	-	
11.	25	65	41	33	34	41	45	55	42	47	-

median r=.3

Subordinates (N=606)

1	2	3	4	5	6	7	8	9	10	11
-										
21	-									
32	34	-								
29	28	41	-							
19	46	38	38	-						
39	33	49	36	37	-					
31	39	36	37	35	42	-				
18	55	33	32	45	40	40	-			
27	39	28	30	40	37	36	51	-		
31	27	37	37	32	44	43	40	41	-	
19	66	32	29	49	33	42	57	40	32	-

median r=.37

Peers (N=470)

	1	2	3	4	5	6	7	8	9	10	11
1.	-										
2.	18	-									
3.	33	30	-								
4.	37	25	47	-							
5.	22	38	34	37	-						
6.	38	27	28	38	38	-					
7.	28	29	26	34	21	41	-				
8.	14	51	38	30	38	35	30	-			
9.	17	50	20	24	37	25	29	43	-		
10.	37	30	37	38	30	34	39	44	34	-	
11.	13	60	32	29	38	29	28	54	42	30	-

median r=.34

: All decimals omitted.

Table 4 Tests of Significant Differences for Leniency, Range Restriction and Halo Among Four Rater Sources Using Wilcoxon Matched-Pairs Signed-Ranks Test

Index	Comparison Results	Significance Level	
Leniency (means)			
	Self = Superior	T= 14.5	n.s.
	Self = Subordinates	T= 28.5	n.s.
	Self = Peers	T= 22.0	n.s.
	Superior = Subordinates	T= 22.0	n.s.
	Superior = Peer	T= 29.0	n.s.
	Subordinate = Peer	T= 12.5	n.s.
Range Restriction (standard deviations)			
	Self = Superior	T= 15.0	n.s.
	Self = Subordinates	T= 27.0	n.s.
	Self = Peer	T= 30.0	n.s.
	Superior > Subordinate	T= 7.0	p<.05
	Superior > Peer	T= 8.0	p<.05
	Subordinate < Peer	T= 7.5	p<.05
Halo (heterotrait-monomethod correlations)			
	Self < Superior	T= 66.0	p<.001
	Self < Subordinate	T= 39.0	p<.001
	Self < Peer	T= 96.7	p<.001
	Superior = Subordinates	T=706.5	n.s.
	Superior > Peer	T=290.5	p<.01
	Subordinate > Peer	T=260.0	p<.001

Factor Analysis of Performance Ratings by Four Rater Sources

	Self				Superiors				Subordinates				Pe	
	Raw		Residual		Raw		Residual		Raw		Residual		Raw	
	I	II	I	II	I	II	I	II	I	II	I	II	I	II
entative	63	10	62	08	47	32	53	06	53	10	50	07	48	35
	21	40	15	40	66	-30	05	50	22	65	12	36	61	-41
n	59	17	57	16	58	25	53	11	58	27	57	17	59	27
onitor	47	17	44	16	53	33	56	05	51	29	49	20	61	27
Dissem.	15	44	12	43	49	03	29	22	33	54	29	33	56	-07
person	62	18	62	16	69	43	77	08	64	30	62	25	60	21
reneur	41	18	40	16	68	-02	41	42	49	39	45	23	52	10
Handler	18	49	19	49	66	-43	01	66	23	77	12	72	65	-28
ce Alloc.	09	64	11	63	59	-33	08	60	34	53	27	47	56	-35
ator	50	26	51	24	68	-14	29	49	54	33	50	28	62	-09
e before n	3.14	1.29	2.98	1.33	4.22	1.37	2.9	1.64	4.26	1.08	3.24	1.16	3.98	1.23
before on	31.4	12.9	29.9	13.3	42.2	13.7	29.6	16.4	42.6	10.8	32.4	11.6	39.8	12.3
ance explained	44.3		43.2		55.9		46.0		53.5		44.0		52.1	

est loadings underlined. All decimals ommitted on the factor loadings.

Table 6 Inter-rater and Inter-source Agreement on Performance Ratings

e	Inter-Source						Inter-Rater		
	Self-Sup (N=270)	Self-Sub (N=306)	Self-Peer (N=282)	Sup-Sub (N=257)	Sup-Peer (N=247)	Sub-Peer (N=274)	Sub1-Sub2 (N=253)	Peer1-Peer2 (N=167)	Me
sentative	.27***	.35***	.28***	.20**	.22***	.22***	.17**	.25***	.
r	.16**	.17**	.14*	.12*	.28***	.11	.14*	.15*	.
on	.18**	.14**	.19**	.19**	.26***	.15*	.19**	.14	.
onment Monitor	-.01	.05	.07	.11	.14*	.06	.06	.05	.
mation Dissem.	.04	.13*	.01	.16**	.02	.13*	.17**	.05	.
sperson	.24***	.13*	.14*	.25***	.22***	.18**	.16**	.12	.
preneur	.13*	.10	.19**	.24***	.16**	.12*	.14*	.03	.
s Handler	.08	.08	.04	.09	.23***	.18**	.02	.20**	.
rce Allocator	.15*	.11	.11	.10	.23***	.12*	.21***	.07	.
iator	.21***	.17**	.18**	.16**	.28***	.23***	.05	.06	.
tational fectiveness	.28***	.13*	.24***	.16**	.35***	.27***	.19**	.24**	.
n r	.16	.13	.14	.16	.23	.15	.16	.12	

<.05
<.01
<.001

Table 7 Regression of Dimensional Performance Ratings on Formal Performance Appraisal[1]

formance Dimension	Self (1)	Self (2)	Superiors (1)	Superiors (2)	Subordinates (1)	Subordinates (2)	Peers (1)	Peers (2)
Representative	-.00	-.01	-.02	.14	-.04	.05	.03	.06
Leader	.03	.07	.14	.16	-.07	-.07	-.04	-.00
Liaison	.00	.06	-.03	-.12	.06	-.02	.05	-.04
Environment Monitor	-.12*	-.10	-.05	-.22**	.00	.00	-.08	-.04
Information Disseminator	.00	.07	-.07	.01	.07	.08	-.03	-.02
Spokesperson	.02	.04	.13	.17	.06	.05	.07	.01
Entrepreneur	.13*	.01	.12	.03	.08	.04	.03	.05
Crisis Handler	.02	.08	.02	-.06	.06	-.01	.06	.10
Resource Allocator	.11	.04	.01	-.00	-.09	-.03	.05	.04
Negotiator	.03	.00	-.00	.09	.08	.04	.07	.04
Expectational Effectiveness	.13*	.17*	.30**	.24**	-.01	.04	.20**	.15*
F	2.00*	1.97*	7.69**	4.93**	2.34**	.90	4.27**	2.41**
df	11,317	11,238	11,247	11,189	11,531	11,398	11,434	11,341
R^2	.07	.08	.26	.22	.05	.02	.10	.08

[1] Two formal performance appraisal ratings used as dependent variable, (1) is the first rating obtained at the time the demensional ratings were collected, (2) is the second rating obtained eighteen months later.

* p <.05
** p <.01

Appendix

Managerial Role Descriptions

Not at all	To a very slight extent	To a small extent	To a moderate extent	To a considerable extent	To a great extent	To an extreme extent
1	2	3	4	5	6	7

1. Organizational Representation Role - Involves the manager in a variety of symbolic, social and ceremonial activities. These obligations often arise because other people insist on involving the manager due to his/her formal authority and status in the organization. Examples include speaking at employee luncheons, participating in civic affairs, signing contracts with key customers, and presenting certificates of course completions to employees.
2. Leader Role - Concerns the manager's overall efforts to motivate and develop his/her subordinates to perform the necessary work. The manager tries to create an appropriate atmosphere and to establish good interpersonal relations with subordinates. Examples include hiring new employees, training and coaching subordinates, conducting performance appraisal interviews, and reprimanding subordinates for poor performance on projects.
3. Liaison Role - Sees the manager maintaining a network of contacts and information sources outside his/her organizational unit. This includes relationships with people elsewhere within the organization (peers and other higher ups) and with contacts outside the organization. Examples include contacts with peer managers, company committee work, industry professional meetings, and social events with key outside contacts or internal contacts.
4. Environment Monitor Role - Involves the seeking and receiving of information to better understand the organization and its environment. The manager processes information to be informed, identify problems and understand the changing environment. Examples include skimming memos for relevant information, asking other managers about top management decisions, trading insights with industry contacts, and reading trade magazines.
5. Information Disseminator Role - Sees the manager transmitting information received from outsiders and other subordinates to the appropriate subordinate. The manager may also share accumulated relevant information. Examples include forwarding relevant written data, briefing a subordinate on the background of a new assignment, introducing the subordinate to people with important information, and passing along informal information or news from the informal networks.
6. Spokesperson Role - Involves the manager transmitting information to people outside his/her unit. It may be given to other people within the overall organization, often higher level managers, and to various public groups. Examples include reviewing the unit's results with top management, explaining his/her unit's operation to peer managers, lobbying with outside contacts, and answering questions about company plans at a community meeting and/or internal cross-organizational groups.
7. Entrepreneur Role - Sees the manager searching the organization and its environment to identify opportunities and nonpressing problems to exploit. He/she also initiates and supervises the projects to bring about needed change. Examples include working on creative projects, proposing major changes in the unit's workflow , reviewing a subordinate's progress on a special project, and scanning the environment for opportunities to exploit.
8. Crisis Handler Role - Finds the manager taking corrective action when he/she faces important, unexpected problems or crises. The manager acts because the pressure on his/her organizational unit is too great to ignore. Examples include working on a hot project for the boss, supervising crash programs to solve schedule delays, handling a key customer complaint, and resolving a dispute between two subordinates working on the same project.
9. Resource Allocator Role - Involves the manager allocating organizational resources. This includes approving various authorizations, programming of subordinates' work, and scheduling his/her own time for various activities. Examples include signing purchase requisitions, assigning a subordinate to work on a new project, blocking out a morning for a project status meeting, and readjusting priorities of previously assigned tasks.
10. Negotiator Role - Sees the manager representing his/her organizational unit or the overall organization at various nonroutine negotiations. They can be with other organizations, outside individuals, or with other units within the overall organization. Examples include negotiating a sales contract, making changes in delivery schedules, resolving workflow with other organizational units, and determining payment schedules for contracts.