AN INVESTIGATION OF A METHOD FOR VALIDATING INDIVIDUAL RATERS OF PERFORMANCE AND ITS IMPLICATIONS FOR A GENERALIZED RATING ABILITY

by

Jamie J. Carlyle

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

APPROVED:

H. J. Bernardin, Chairperson

B. v. H. Gilmer

J. A. Sgro

D. A. Bownas

J. J. Hoover

May, 1982

Blacksburg, Virginia

Acknowledgements

I would like to thank John Bernardin for his support, words of wisdom, and unfailing sense of humor throughout the course of this project and my crazy years. Also, I would like to thank my committee members, Drs. Gilmer, Sgro, Bownas, and Hoover for their guidance and support on this project.

To my friends who worked with me on various phases of this project I would like to express my appreciation, especially Teri Ellison, Steve Cecil, Judy George, Joe Orban, and Barry and Robin Riegelhaupt. Finally, I would like to thank my family for their love, understanding and ever-constant sideline cheers--- the end of the tunnel--- FINALLY!

Table of Contents

Page

APPENDICES page

List of Tables

Introduction

Assessment of employee performance is critical in any type of organization. Insofar as this assessment or appraisal of performance is used as a basis for making various personnel decisions, the evaluations need to be as accurate as possible. The types of personnel decisions which might utilize performance appraisals include promotional decisions, the identification of training needs, and decisions related to employee development (i.e., serving as performance feedback for the employee, enabling the employee to recognize strengths and weaknesses in performance, and to develop strategies for improving performance in the future). It is essential that resultant performance appraisals in use in an organization demonstrate an acceptable level of validity. Validity of the use of performance information should be demonstrated, regardless of the method used to arrive at this information.

Traditionally, the most commonly used method for appraising employee performance has been that of performance ratings. The extensive use of ratings in organizations is reflected by Landy and Trumbo's (1979) report that 72% of the validation studies in the *Journal of Applied Psychology* since 1955 used ratings as the primary criterion. Though one might expect that more objective, quantifiable measures (e.g., absenteeism, production figures) would be more desirable to use as criteria of effective/ineffective performance, ratings are nonetheless used to a much greater extent. One obvious reason for this is that these "objective" indices simply do not exist for all jobs and when they *do* exist, they are often not comprehensive representations of work performance. Also , as several authors have noted, indices of this type are often strongly influenced by factors beyond the employee's control (e.g., Cascio & Valenzi, 1978). Appraisal methods other than ratings, such as personnel comparison techniques, often do not provide the ratee with specific behavioral feedback and create animosity among those being compared. For these reasons, ratings seem to be the method of preference. Generally speaking, the validity of ratings refers to the extent to which the rating reflects the actual performance demonstrated by the individual being evaluated. That is, one is inferring the actual demonstrated performance level from the rating given. Validity, by definition, is concerned with inferences made about the scores (ratings) as opposed to the scores (ratings) themselves.

One of the purposes of the present study is to consider the various strategies for validating ratings, and to demonstrate a new validation strategy which attempts to validate raters as opposed to ratings. The emphasis, then, will be on the individual rater. Appendix A contains a complete review of the various attributes of the rater, and the ratee, and of the parameters of the rating process which affect the validity and important psychometric characteristics of resultant ratings. The section to follow will examine the various approaches for establishing the validity of ratings that have been used in the past, as well as explore the possibility of validating individual raters as opposed to ratings. The possibility of validating individual raters also leads into a discussion of the notion of the existence of a generalized rating ability, and how this issue might be investigated in the context of previously described validation research.

Review of the Literature

Validation Strategies

One strategy that has been used to validate ratings is that of correlating the ratings with other criteria of performance (i.e.,criterion related validation strategies). As noted previously, however, ratings are often used to evaluate performance.Because no other measures of performance data (i.e., readily verifiable data, such as productivity records) are available for validation purposes, researchers have to rely on other ratings to serve as criteria for validation purposes.

Some attempts have been made to validate ratings against "objective" performance data. Although not all of the studies cited below were specifically concerned with validating ratings, they do examine correlations between ratings and various objective criteria of performance.

The most frequently used objective criteria for validation purposes seem to be absenteeism and turnover, probably because these data are easily obtained in most organizations. Several studies have shown modest correlations between performance ratings and absenteeism and/or turnover (Latham & Wexley, 1977: Ronan & Latham, 1974; Seashore, Indik & Georgopoulos, 1960). Productivity data have also been correlated with performance ratings in several studies (e.g., Latham & Wexley, 1977; Ronan & Latham, 1974; Severin, 1952) as well as indices of productivity such as sales performance, new accounts, number of arrests, etc. (Cascio & Valenzi, 1978; Waters & Waters, 1970). Other objective performance measures that have been correlated with performance ratings have included such criteria as training records (e.g., Severin, 1952), grade point average (e.g., Miner, 1917), and various intelligence and performance test scores (Bayroff, Haggerty & Rundquist, 1954; Bell, Hoff & Hoyt, 1963; Hausman & Strupp, 1955; Whitla & Tirrell, 1953).

As noted by Kavanaugh (1971), however, a major problem in trying to validate ratings against ojbective data (or even non-objective data, such as other ratings) is that it must be assumed that those other criteria are valid. Some of the data used as criteria in the studies cited above are questionable in this respect. The process of validating ratings can thus become a problem of infinite regress (i.e., ratings are validated against production records, which are validated against absenteeism records, etc.). As has been noted previously, it is often difficult to obtain any objective, valid indices of performance to serve as criteria in validating ratings. For these reasons, other methods have been sought to demonstrate the validity of performance ratings.

Kane and Lawler (1979) have suggested that construct validation is the "only relevant type of validity to consider since the other major type -- criterion-related validity -- requires the availability of a more nearly ultimate measure of job success," (p. 427). One method which has proven useful in the demonstration of construct validity is the application of the Multitrait-Multimethod matrix analysis (MTMM) to ratings of performance. This technique, formally introduced by Campbell and Fiske (1959), requires the assessment of two or more traits by two or more methods These assessments are then used to set up a correlation matrix such that all possible correlations are computed among the scores obtained when all the traits are measured by all the methods. Validity is then assessed by examining the correlations, seeking evidence for convergent and discriminant validity. Convergent validity is evidenced by significant correlations between different methods on the same traits (e.g., significant correlations in the validity diagonals of the matrix). Evidence for discriminant validity is threefold (Kavanagh, MacKinney & Wolins, 1971). That is, not only should the correlations in the validity diagonals be higher than the correlations in the same columns and rows (i.e., where neither trait nor method are in common), but also the correlations in the validity diagonal must be higher than the correlations between that trait and all other traits assessed by that particular method. Finally, the "pattern of trait interrelationships should be the same" both within and between methods used to assess those traits (Kavanagh et al., p. 35). In other words, convergent validity represents the extent to which methods differ in their measurement of the same traits, while discriminant validity represents the extent to which methods differ in their measurement of different traits.

This MTMM matrix analysis can be applied to performance ratings by considering different raters (or different rating instruments) as the Methods and ratings of performance as the Traits. This approach has been applied to performance ratings by a number of researchers (Charest, Cowart & Goodman, 1969; Dickinson & Tice, 1973; Goodman, Furcon & Ross, 1969; Ivancevich, 1977; Kavanagh et al., 1971; Lawler, 1967; Tucker, Cline & Schmidt, 1967; Zedeck & Baker, 1972). Generally, these studies have shown moderate evidence of convergent validity, but only limited evidence of discriminant validity. Questions have been raised, however, as to whether one should be concerned with the degree to which convergent validity is present or the degree to which it is absent. Kavanagh et al. (1971) believe the appropriate approach to take is to assess the degree to which convergent validity is present as evidenced by a significant ratee main effect (i.e., performing an analysis of variance of appraisal scores "attributable to source (raters), object (ratees) and performance dimension factors and their interaction") (Kane & Lawler, 1979, p. 427). The size of the monotrait-heteromethod correlations (i.e., those in the validity diagonals) is directly proportional to the strength of the ratee main effect, and inversely proportional to the strength of the rater by ratee interaction.

Several researchers have cited a number of problems with the MTMM matrix approach. Although not specifically addressing correlations in an MTMM matrix, Guilford (1954) noted that constant errors (e.g., halo, leniency, contrast effect) can bias correlations such as those found in an MTMM matrix. Adjustments would therefore need to be made in order to ensure correlations are not biased in this way. Eliminating these sources of error variation should increase reliability accordingly, and the possiblity of higher reliability would also increase the possibility of higher validity. Lawler (1967) has noted that one could conceivably demonstrate the convergent and discriminant validity of ratings, and still find that the ratings were not actually valid measures of the performance dimension one intended to measure. For instance, it might be the case that two sets of raters (e.g., peers and supervisors) agree perfectly in their ratings, providing strong evidence of both convergent and discriminant validity. Yet, it could be that both of these sets of raters are incorrectly observing behavior (thus biasing the ratings) in exactly the same way. Therefore, as highlighted above, there are some potentially serious problems associated with the MTMM matrix technique. In light of problems such as these, some researchers have opted for the approach of inferential validity for validating ratings. That is, by demonstrating that ratings are psychometrically sound (i.e., reliable, unbiased), one might infer the validity of these ratings from this evidence. Psychometric qualities and accuracy might serve as evidence for inferring the validity of ratings. Each of these will be discussed below.

Rating efforts have long been a concern of researchers relying on data gathered from ratings made by individuals. Constant rating errors, which are systematic distortions of the ratings made by raters, have received much attention in the literature on performance appraisal. The constant errors most frequently studied relevant to performance appraisal ratings include halo, leniency/severity and central tendency. Generally halo has been regarded as the failure of the rater to distinguish between different dimensions when rating an individual's performance, thus resulting in spuriously high intercorrelations among performance dimensions. Another type of constant error is restriction of range of the rating scale. One such restriction is leniency/severity, the tendency to use the extreme ends of the rating scale, thus not discriminating among performance dimensions. Another restriction, central tendency, is found when the rater utilizes only the central points of the rating scale when rating individuals. Due to this limited use of the scale values, a lack of discriminability occurs, thus rendering the ratings useless for decision-making purposes (Cascio, 1978).

Other types of errors which affect ratings include those that are due primarily to conscious distortions of the ratings by the individual rater. An example of this type of error is logical error, which is the error resulting from raters assuming what seem (to them) to be logical relationships among dimensions, and assigning ratings accordingly. This results in spurious intercorrelations among performance dimensions.

Similarly, proximity errors result in spurious intercorrelations, but these errors result from the rater's tendency to assign similar ratings to dimensions close to one another on the rating scale. The problem that has arisen, however, in using these indices as evidence for the validity of ratings concerns the assessment of these indices. As Saal, Downey and Lahey (1980) noted, there has been considerable disagreement as to how these indices should be measured. Given the inconsistencies in measuring these criteria, it is difficult to make generalizations concerning these criteria from study to study (e.g., "halo" in Study A is entirely different from "halo" in Study B).

Other problems with using these psychometric criteria of ratings as evidence of validity have been brought to light by research conducted in an attempt to eliminate these rating errors. Using videotapes of performance Borman (1975) attempted to train raters to reduce halo errors. Although successful in reducing halo, this training decreased interrater reliability (which, as will be discussed later, is also used as a source of evidence for the validity of ratings). In a later study, Borman (1979), using video-tapes of performance as in his earlier study was able to significantly reduce halo (in some cases) by training but was not able to improve accuracy. Finally, Bernardin and Pence (1980) were able to decrease halo and leniency via rater training, but at the same time decreased the accuracy of the resultant ratings.

Because of the aforementioned problems in assessing these psychometric criteria of ratings, other psychometric criteria have been utilized as evidence for inferential validity. Interrater reliability is one such criterion. Although the assessment of interrater reliability is part of the MTMM matrix analysis discussed previously, it may also be examined without such a matrix analysis. Interrater reliability is the "extent to which two or more raters independently provide similar ratings on given aspects of the same individual's behavior" (Saal, Downey, & Lahey, 1980). Several authors have questioned whether interrater reliability is an appropriate criterion for validity, as it is conceivable that raters may be systematically biasing their ratings while still achieving high interrater reliability (e.g., Freeberg, 1969; Wherry, 1952). For example, it may be the case that two sexist male raters are evaluating the performance of a female employee. The ratings they give this employee may be in perfect agreement, yet invalid because they have been biased by the raters' prejudiced attitudes toward female employees (e.g., a behavior that they see as "decisive" in a male employee is viewed as "stubborn" in a female employee). In addition, interrater reliability seems to be plagued by the same problems previously discussed concerning constant rating errors, noteably inconsistencies in definition, measurement, and research methodology (Saal et al., 1980). However, interrater reliability could be very useful for inferring validity if used with other sources of information which might aid in determining whether such biases are operating (e.g., accuracy measures).

This final source of evidence for inferential validity, rating accuracy, may be defined as the extent to which a rater's ratings approximate the "true" or exact measure of an individual's performance. Of course, the obvious problem with accuracy lies in its measurement, i.e., if one knew what the "true" performance of an individual were in the first place, ratings would not be necessary. As Borman, Hough and

Dunnette (1978) noted, one way of dealing with this problem is to take several different measures of the construct for which one is developing the "true score" and intercorrelate these measures. If these measures demonstrate convergent and discriminant validity (e.g., Campbell & Fiske, 1959), then the mean of the scores may serve as a true score against which one may evaluate the accuracy of a particular rater's ratings. A comparison of this sort, then, would involve much the same procedure as a criterion-related validity assessment. That is, one is comparing performance ratings to other measures of performance (i.e., "true scores" which should be somewhat more objective than "one-shot" ratings).

True scores developed in this way might be viewed as the "ultimate criterion" of performance. As defined by Thorndike (1949) the "ultimate critierion is the complete final goal of a particular type of selection or training.... A criterion is ultimate in the sense that we cannot look beyond it for any higher or further standard in terms of which to judge the outcomes of a particular personnel program." (p. 121). Thus, if performance is carried out to reflect certain predetermined levels of effectiveness, then these predetermined levels (true scores) would represent the "ultimate criterion" for validating ratings. Unfortunately, since such true scores do not typically exist in the "real world" there exists a need for the development of some method/measures which are essentially equivalent to these true scores (i.e., a replacement for the true score method). Basically, what this would involve is the demonstration of the validity of the scores from the new method (i.e., validating the criterion) by means of the true scores which purport to measure the same constructs. One of the purposes of the present study is to demonstrate the validity of such a method in order to propose a replacement method for deriving criteria used to validate raters.

Assuming one has a suitable criterion for validation purposes, there is an additional problem in that there has been considerable disagreement as to how accuracy should be computed. For example, Cronbach (1955) demonstrated that the D^2 statistic commonly used to measure accuracy actually contains four separate components of accuracy: elevation, differential elevation, stereotype accuracy, and differential accuracy. These different components were later shown to have very low intercorrelations (Cline, 1964). Several authors have concluded that the most conceptually appropriate component of accuracy to use in assessing a rater's accuracy is the differential accuracy component (e.g., Borman, 1979; Hastorf, Schneider, & Polefka, 1970; Sechrest & Jackson, 1961). The use of only this index of accuracy would eliminate the response bias components in accuracy measurement (i.e., elevation, differential elevation, and stereotype accuracy) (Hastorf, Schneider, & Polefka, 1970).

Assessing Accuracy

It would seem from the literature presented above that accuracy may be a very appropriate measure to serve as evidence for inferring the validity of performance ratings. As Borman (1979) has noted, accuracy is the "critical criterion for judging quality of performance ratings" (p. 412). It must be noted, however, that while ratings may be accurate as determined by the computation of differential accuracy scores (Cronbach, 1955) they may still be invalid (in terms of an analysis of variance approach for obtaining evidence of validity as proposed by Kane and Lawler, 1979). While accuracy assessment yields a specific statistic, validity assessment yields no such single statistic. Instead of being a one-shot approach, validity assessment requires different methods and procedures that produce evidence of the fairness (or appropriateness) of the hypotheses generated by the ratings (i.e., that rating reflect actual performance). While accuracy can serve as part of this evidence, accuracy alone is not equivalent to validity. As noted by Campbell (1976) the demonstration of construct validity, in particular, requires several different sources of evidence since with construct validity one is attempting to obtain "a measure of a characteristic which is deemed important by somebody but for which there is no already available indicator" (p. 203). One source of evidence suggested by Campbell to contribute to the establishment of construct validity is the demonstrated correlation of measures or variables which have been hypothesized to measure the same thing. Accuracy scores (i.e., correlations of ratings to true scores) would provide such evidence. Thus, though accuracy may not be sufficient for the demonstration of construct validity, it can serve as one piece of evidence for it.

Nonetheless, assuming one wishes to use accuracy as a source of evidence of validity, the problem then becomes one of determining the most fruitful approach for examining rater accuracy. Two of the first studies published which looked specifically at rater accuracy (using the differential accuracy statistic) used a series of videotaped performance vignettes (Borman, 1975; 1979). In these studies, intended true scores were established for two different jobs. A total of sixteen videotaped vignettes were developed and subsequently rated by experts for the purpose of establishing true scores of performance. As mentioned previously, the means of these expert ratings were used as the final true scores, which were correlated with the ratings given by subjects (raters) in these studies. These correlations served as accuracy scores for the raters. This method is particularly useful because it enables researchers to look at rater accuracy in a controlled setting. However, the one-shot approach used by Borman (1975; 1979) does not allow for any conclusions to be drawn concerning raters' accuracy in assessing performance distributions of the ratees (i.e., individuals' performance over time), which would seem to be more representative of "real world" rating situations.

Another approach for examining rater accuracy is that of immediate scoring by the raters, e.g., having the raters rate behavior in some way as it occurs and compare these ratings to later summary ratings made by those same raters. A method that might be utilized with this approach is that of diary keeping (Bernardin & Walter, 1977). Specifically, raters would be asked to record critical incidents of ratee behavior throughout the performance appraisal period. These behaviors could be rated by the raters on previously established performance dimensions each day (or week), and then the raters could do a summary rating at the end of the performance appraisal period on those same dimensions. Differential accuracy scores for the raters could then be computed by comparing the two sets of ratings for each rater (Bernardin, 1979).

A potential problem with the method, however, is that rater biases could be operating which influence both the ratings done each day (week) and the summary ratings done at the end of the appraisal period. To dissuade the effects of this type of bias, raters might instead simply record the critical behaviors (making no ratings until the final summary ratings) and then submit these incidents to another group of informal raters for rating. Differential accuracy could then be assessed by comparing these ratings to the raters' summary ratings.

The differences in the two approaches which utilize the diary keeping technique discussed above serve to highlight an important distinction in the assessment of rating -- that between validity and reliability. As Campbell and Fiske (1959) noted "Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods" (page 93). If the same raters assess behavior both periodically (e.g., by recording critical incidents) and at the end of the appraisal period (e.g., performance ratings), one can assess the degree of reliability in the ratings (since the critical incidents and the performance ratings may be viewed as two measures of behavior completed by the same rater). If different raters assess the behavior of the same ratees (e.g., critical incidents written by one group of raters compared to the ratings by another group of raters observing the same ratees), the validity of the ratings can be assessed (i.e., construct validity, since the two assessments are made by different raters looking at the same ratees). It is interesting to note that most organizations today typically have only one rater observe an employee's performance, in contrast with the validation procedure described above which requires using multiple raters.

Constraints of the organization often prohibit the use of multiple raters, unfortunately. However, if it can be demonstrated that the two methods yield comparable results, then the method utilizing only one rater could be justifiably substituted for the so-called validation method. One of the purposes of the present study is to determine the comparability of the two methods.

The comparison of two such methods has yet to be done in a controlled setting such as that described by Borman (1975; 1979). If the two methods are shown to yield comparable results, then this method would seem to be a valid one. Nonetheless, the argument might be made that validating a rater with this particular method in a given situation says nothing about the validity of the rater in other situations in which s/he would have to rate employees. The need exists, then, to demonstrate that a rater who makes valid ratings in one setting can also make valid ratings in another setting. In other words, it must be determined that rating accuracy is a generalized ability.

Rating as a Generalized Ability

There is evidence from the literature on person perception that raters may differ in their ability to make accurate judgments about other people. In his review of the literature examining individuals' abilities to accurately judge others' traits and/or emotions, Taft (1955) concluded that certain factors seem to be correlated with this ability. The most notable of these factors are intelligence, training in Psychology, dramatic and artistic interests, emotional stability, and social skills. However, because of methodological problems with earlier studies done in this area, some researchers have concluded that these studies have provided little evidence for a generalized ability to rate others accurately. Instead, they suggest that various response sets may be stable over time, as opposed to an accuracy ability (e.g., Bronfenbrenner, Harding, & Gallwey, 1958; Crow & Hammond, 1957; Sechrest & Jackson, 1961). Nonetheless, studies which have examined raters' accuracy in judging others' performance have provided limited evidence that this accuracy may be a generalized ability (i.e., persons who are accurate in rating certain performance dimensions are also accurate in rating other performance dimensions) (Borman, 1979; Mullins & Force, 1962). It has been suggested that instead of being a unitary process or ability that is responsible for a particular rater's accuracy in judging others, there may be many related abilities and processes which converge in some way to enable an individual to make accurate ratings (Taguiri, 1969).

As noted previously, a comparison of the two methods (i.e., the "reliability" method and the "validity" method) has yet to be done in a controlled setting such as the one described by Borman (1975; 1979). Borman's studies are controlled in the sense that raters are viewing videotapes of managers and college recruiters who are performing at predetermined levels of effectiveness, in one particular setting (i.e., each manager and each recruiter is seen in a brief interview situation only once). The availability of true scores in such a setting makes possible the comparison of the two validation methods (i.e., the true scores can serve as the ultimate criterion against which to compare the results of these two methods). The reliability method can also be used in another type of setting as well as this controlled setting (i.e., controlled setting as described by Borman). The other setting which will be used for comparison purposes in the present study will be a college classroom, with raters recording critical incidents of their instructors (over a four week period) and rating their instructors' performance. By comparing estimates of accuracy derived in this setting with accuracy estimates derived using the Borman videotapes, it will be possible to assess evidence for a generalized rating ability that may exist.

As mentioned earlier, Appendix A contains a complete review of the rating process and individual differences variables that might affect this rating process. One individual difference variable deserves mention here, nonetheless. This variable is cognitive complexity. As defined by Schneier (1977) cognitive complexity is the "degree to which a person possesses the ability to perceive behavior in a multidimensional manner" (p. 541). As such, it may be seen as a fairly stable personality trait which could potentially affect an individual's ratings of others' performance. Given the types of individual differences variables studied in the literature on person perception, cognitive complexity is one variable which intuitively seems to relate well to the notion of a generalized rating ability.

A number of researchers have noted the relationship between cognitive complexity and performance ratings, as exemplified by the statement made by Jacobs, Kafrey, & Zedeck (1980), "...cognitive complexity is one property of the rater which relates to effective performance evaluation..." (p. 634). In a frequently cited study of cognitive complexity, Schneier (1977) found cognitive complexity to be significantly related to performance ratings. Specifically, Schneier found that cognitively complex raters exhibited less leniency error, less halo error, and less restriction of range error than cognitively simple raters, when using a Behavioral Expectation Scales format. Furthermore, cognitively complex raters preferred the BES format, while the cognitively simple raters preferred to use a simpler format. Other studies which have examined the relationship between cognitive complexity and performance ratings have not found significant relationships (e.g., Borman, 1979). The present study will attempt to shed new light on this issue by examining the effects of cognitive complexity on rating accuracy in the context of the generalizability of rater validity. This is done by averaging accuracy estimates derived from two rating situations (i.e., rating the performance of managers viewed on the Borman tapes and instructors viewed in a classroom situation), for each rater. These averaged accuracy estimates can serve as indices of a "generalized rating ability." Given these indices of generalized rating ability, the effects of individual raters' cognitive complexity on the accuracy of their ratings can be examined.

Summary of the Present Study

In summary, the present study examines an approach for validating individual raters as opposed to ratings, using a measure of rater accuracy. As noted earlier in this discussion, there seems to be great potential benefit in identifying valid raters, since research has shown that the rater accounts for much of the variance found in ratings (Landy & Farr, 1980). The validation strategies used in this study have a number of steps. Each of these steps will be outlined below.

Raters involved in this study maintained diaries (i.e., recorded critical incidents of ratees' behavior) based on their observations of ratees' performance. Raters observed ratees whose performance was displayed in a single, controlled setting (i.e., Borman videotapes) and ratees in a "real world" setting whose performance is viewed over a period of time (i.e., college instructors). Raters rated both sets of the observed ratees. The incidents which were recorded by the raters were typed and made anonymous (i.e., incidents were not identified with their authors or the ratees about whom they were written). These incidents were grouped according to the performance dimension on which they were written, randomized within these groups and then submitted to Critical Incident Raters (CIRs) for review. Critical Incident Raters rated the dimensions on their effectiveness in relation to the context (job) for which they were written. Mean effectiveness values were derived for each incident and linked back to the ratees for whom each incident was written, and the raters who wrote them. A mean score was derived for each ratee on each dimension (based on the mean effectiveness ratings given by the CIRs to individual raters' incidents). The mean ratings of effectiveness for each ratee (on each dimension) was correlated with the actual ratings assigned to each ratee by the individual raters. This provided a reliability estimate for each rater and for each dimension within raters (since it used two forms of the same rater's observations, i.e., critical incidents and performance ratings by the same rater on the same ratee). This estimate, which was based on the rater's accuracy (since the estimates based on the Borman videotapes were comparable to the Borman true scores), also served as evidence of the rater's validity (inferring validity from accuracy). A comparison of summary ratings made by each rater with the mean critical incident ratings written by other raters was also made for the ratees on the Borman tapes, which served as an estimate of the rater's validity (i.e., the critical incidents and ratings came from two different sources).

Being able to identify valid raters in this way will be useful for learning more about the process of making performance evaluations. If valid raters can be identified in this way, they can be compared to less valid raters, and other comparisons can be made to determine if these labels of "valid" and "less valid" raters are sustained across rating situations, shedding light on the notion of a generalized rating ability. Furthermore, investigations of the relationship between cognitive complexity and generalized rating ability can be conducted.

In light of the discussion above, the following hypotheses can be made concerning the present study:

1) Mean values of critical incidents scaled by Critical Incident Raters and written by observers will be significantly correlated with the true scores derived by Borman for the performance of videotaped managers.

2) Mean values of critical incidents scaled by Critical Incident Raters and written by <u>other</u> observers will be significantly correlated with individual raters' ratings. This estimate of validity will serve as a data base for inferring accuracy.

3) Mean values of incidents scaled by Critical Incident Raters and written by a particular rater will be significantly correlated with ratings made by that same particular rater (this prediction applies to both ratee samples, i.e., managers and instructors). This estimate reliability will serve as a data base for inferring accuracy.

4) There will be a significant correlation in the rank ordering of rater accuracy estimates (i.e., correlations between ratings and critical incidents scaled by Critical Incident Raters) derived in Hypothesis 2 and Hypothesis 3.

5) The rater accuracy estimates derived by correlating Borman true scores with ratings of managers will be significantly positively correlated with cognitive complexity scores of the raters (e.g., the more cognitively complex the rater, the higher the accuracy estimate).

   5a) Rating accuracy estimates for the instructor ratings as derived in Hypothesis 3 will be significantly positively correlated with cognitive complexity scores of the raters.

6) Accuracy estimates for the managers' ratings (as defined by correlations between ratings of managers and Borman true scores) will be significantly correlated with accuracy estimates for instructor ratings (as defined by correlations between critical incidents and ratings for instructors).

7) The two sets of accuracy estimates described in Hypothesis 6 will be averaged for each rater, and will be significantly correlated with rater cognitive complexity.

Method

Subjects

Undergraduate students at Virginia Polytechnic Institute and State University participated in the present study. Twenty-nine subjects participated in the study for a total of four weeks, while 38 subjects participated in one one-hour session. All subjects received appropriate credit points for their participation.

Materials

Manager videotapes - A series of videotapes, developed by Borman (1977) were used in the present study. The videotapes depicted a number of managers, each interviewing the same "problem" employee. Each interview lasted from five to seven minutes. Scripts were developed for the interviews so that the managers' performance effectiveness varied on a number of dimensions, according to previously established levels. Rating scales were developed for the manager job using the methodology described by Smith and Kendall (1963) for developing behaviorally anchored rating scales (BARS). True scores of performance were developed for each of the managers on the various dimensions. A complete description of the developement of the videotapes, rating scales, and true scores may be found in Borman (1977).

Instruments

Behaviorally Anchored Rating Scales - Two different BARS were used in the present study. For rating instructors, a ten-point BARS developed in previous studies of faculty evaluation (e.g., Bernardin & Walter, 1977) was used, representing five dimensions of performance. The instructors' BARS is presented in Appendix B.

In order to rate the performance of managers depicted on the videotapes, a seven point BARS, developed by Borman (1977) was used, representing four dimensions of performance.

Summated Scales - Two sets of summated scales were used in rating the critical incidents written about the managers and instructors. The summated scales were used by Critical Incident Raters to rate the incidents on their overall effectiveness in the context of the dimensions for which they were written. One set of summated scales, which used general behavior descriptors of high, medium and low performance on each dimension in the managerial BARS has been developed by Borman (1977). Four descriptors were randomly selected from each of five dimensions to make up the scale. The other summated scale used general behavioral descriptors of high, medium and low performance based on the BARS, but this scale described instructors' performance. Four descriptors were randomly selected from each of five dimensions to make up the scale. The summated scales for managers and instructors are presented in Appendix C.

Individual Differences Measure - The 29 subjects writing critical incidents for managers and instructors were administered a cognitive complexity scale during the five weeks in which they were participating in the study. The Kelly Repertory Grid was the measure of cognitive complexity used in this study (Vannoy, 1965).

Procedure

The first four weeks of the study, the procedure was as follows:

Week 1 - Thirty eight subjects were trained to write critical incidents.

These subjects were given observational diaries on which they were asked to write at least two incidents on each of five dimensions for three different instructors during that week. The diaries which were used are presented in Appendix D. Diaries were turned in to the experimenter at the end of the week for review so that the subjects could be given feedback on the quality of the incidents. This feedback was given only on the incidents recorded during the first week. Also during the first week, subjects viewed one videotape of a recruiter interviewing a job applicant and were asked to write critical incidents of the recruiter's performance on four dimensions. The recruiter tape was developed by Borman (1977) in the same way in which the manager tapes were developed. These incidents were also collected at the end of the session in order for the experimenter to review them and give subjects feedback on the quality of the incidents.

Week 2 - Subjects were given written comments concerning the quality of their critical incidents written about their instructors and the recruiter. At this time, they also received observational diaries for recording subsequent critical incidents on their instructors (i.e., the same instructors they previously observed) following the same procedure in week 1.

Subjects were given the BARS to be used in rating the manager tapes. After becoming familiar with the BARS and receiving a brief lecture concerning their use, subjects observed the performance of one of the managers on the tape, recording critical incidents on each of the four dimensions at the conclusion of the tape. After recording critical incidents, subjects rated the manager on each of the four dimensions using the BARS. This procedure was followed for five more managers on the tape.

Week 3 - Subjects submitted critical incidents written about their instructors during the second week, and received new observational diaries on which to record critical incidents about instructors during the third week.

Week 4 - Subjects followed the same procedure in week 4 as in week 3.

Week 5 - After submitting critical incidents written during the fourth week, subjects were given the BARS for instructors and asked to rate each of the instructors about whom they had written critical incidents on each of the five dimensions.

Critical Incident Rating Procedure - In order to determine the effectiveness levels of each of the critical incidents written for the managers' performance, the following procedure was followed:

1) All critical incidents recorded by the subjects on the manager tapes were collected and typed. Individual incidents were not identified with the rater who wrote them, and any non-behavioral information contained in any of the incidents which might have identified the manager for which the incident was written was omitted from the incident.

2) Incidents were grouped by the four dimensions for which they were written. Within each dimension, incidents were randomly ordered, placed on rating forms, and distributed to an independent group of subjects. These subjects (hereafter referred to as Critical Incident Raters, or CIRs) were trained on the critical incident methodology and familiarized with the summated rating scales previously described. Each of these CIRs rated the overall effectiveness of the critical incidents using the managerial summated rating scales.

To minimize the number of incidents that had to be evaluated by each CIR, the incidents (by dimension) were randomly divided into eleven separate forms (each form containing approximately 200 incidents). Each of the eleven forms were rated by seven different CIRs (11 forms X 7 CIRs per form = 77 CIRs participated in this procedure).

3) Mean effectiveness values were derived for each incident and linked back to the manager for which each was written.

In order to determine the effectiveness levels for each of the critical incidents written about instructors' performance, the same procedure (as described above) was followed, except that the summated scale used by the CIRs was the one developed for instructors. A total of 15 forms (with approximately 200 incidents per form) was used, with seven CIRs rating each form (15 forms X 7 CIRs per form = 105 CIRs participated in this procedure).

Results

Validity of the Incident Rating Procedure

All critical incidents written about the Borman (1977) videotape managers' performance were collected and made anonymous (i.e., any information contained in the incidents which might identify the writer of the incident or ratee about whom the incident was written was removed). The incidents were then grouped and randomized by dimensions. Eight forms were used, each of which consisted of approximately 210 incidents (with all four dimensions being equally represented on each form). Incidents on each of the eight forms were then rated for effectiveness in the context of the dimensions for which they were written by a minimum of seven independent critical incident raters (CIRs), i.e., subjects who had not viewed the Borman tapes. These incidents were then linked to the managers for whom they were written. A mean effectiveness rating was derived from the CIR ratings of the incidents (across all raters) for each of the managers for each of the four dimensions. These mean values were then correlated with the Borman true scores. These correlations, which were computed across managers, were as follows: Structuring and Controlling the Interview, $\underline{r}(36)=.85$, $\underline{p} < .03$; Establishing and Maintaining Rapport, $\underline{r}(36)=.91$, $\underline{p} < .01$; Reacting to Stress, $\underline{r}(36)=.96$, $\underline{p} < .01$; and Obtaining Information, $\underline{r}(36)=.91$, $\underline{p} < .01$. The correlations on the four dimensions were converted to Fisher z scores, averaged and converted back to an overall correlation. This correlation between the critical incidents and the Borman true scores was found to be significant ($\underline{r}(36)=.91$, $\underline{p} < .01$) as predicted by Hypothesis 1. This significant correlation may be seen as an index of the validity of the critical incident recording/rating method for evaluating performance, indicating that the method was a "valid" one for evaluating performance.

One problem with using only the correlation between mean critical incident values and true scores to test Hypothesis 1, however, is that there may be level effects that would go undetected. The two sets of scores may be highly correlated, yet represent two different views of the ratees' performance. That is, the ratees might be be ordered in the same way with the two methods (thus highly correlated) but have scores resting at different ends of the scales. To determine whether this was the case in the present study, an analysis of variance was performed using the two methods (i.e., critical incidents and true scores) as the dependent variable. The analysis of variance performed on these ratings indicated that level effects were not present (with a nonsignificant effect of Method, F(1,15)=.30). Table 1 contains the complete ANOVA summary table for this analysis. As can be seen in Table 1, the interaction of Manager X Dimension and the interaction of Method X Dimension are marginally significant. Post hoc multiple comparison tests of these interactions revealed no discernible patterns, however.

Assessing Rater Accuracy Via a "Validity" Method

An attempt to assess the same performance by two different methods can be viewed as a method which yields a measure of construct validity (Campbell & Fiske, 1959). The critical incidents for the Borman videotaped managers' performance, scaled by the Critical Incident Raters, were used for this purpose in the following way: Each of the incidents written for a manager except the incidents written by a particular rater (e.g., subject #1) were collected by dimensions and a mean score (the CIR ratings on those incidents) was derived for that manager for each

Table 1

Analysis of Variance to
Examine Level Effects

Source	df	SS	MS	F	p
Between Subjects					
Method*	1	.13	.13	.30	.59
Within Subjects					
Manager	5	37.17	7.44	7.99	.02
Manager X Method	5	4.65	.93	2.14	.11
Dimension	3	.05	.02	.01	.99
Dimension X Method	3	3.89	1.30	2.98	.06
Manager X Dimension	15	14.76	.98	2.26	.06
Manager X Dimension X Method	15	6.52	.44		
Total	47	67.17			

*Method is treated as a Subjects term in this analysis. To test for the effect of Method, the Manager X Dimension X Method interaction was used as the error term.

dimension. This procedure was followed for each of the six managers, omitting the same subject's incidents each time, until each of the 38 raters' incidents were omitted once from the mean scores for each of the six managers. Hence, 38 mean scores on each dimension for each of the six managers were derived, each one based on 37 mean incident values (i.e., 38 total subjects' mean incident values - one subject's incidents = 37 total mean incident values). CIR mean scores were then correlated with the ratings given to managers by a particular rater's Correlations were computed across managers for each rater by dimension. The mean score used for this purpose, however, was the one for each manager in which that particular rater's incidents were omitted from the calculations (hence, the different methods are methods which use different sources of observations of the ratee's behavior, i.e., ratings made by a particular rater were correlated with the critical incidents submitted by all other raters). A mean accuracy score was then derived for each rater by averaging across the dimensions.

Hypothesis 2 predicted that the mean values of critical incidents scaled by the Critical Incident Raters and written by other observers would be significantly correlated with individual raters' ratings. The relationship between the ratings made by individual raters and scaled critical incidents written by other raters was found to be significant ($r(36)=.77$, $p < .001$).

Assessing Rater Accuracy Via a "Reliability" Method

Hypothesis 3 predicted that the mean values of critical incidents scaled by CIRs and written by the same observers would be significantly correlated with single raters' ratings. This approach, which attempts to assess the same performance by two similar methods (i.e., sources of observation), can be viewed as a method which yields a measure of reliability (Campbell & Fiske, 1959). The "validity" method described above was used again, except that instead of using critical incidents written by all other observers to compare to each rater's ratings, the critical incidents written by that particular individual being compared were used.

As predicted by Hypothesis 3, there was an overall significant relationship between ratings made by raters and critical incidents written by those same raters ($\underline{r}(36) = .68$, $\underline{p} < .001$).

Accuracy in Rating Instructors

The reliability method as described above was also used to assess the overall relationship between critical incidents written about instructors by each of the subjects and the subjects' ratings of their instructors. Of the original 38 subjects used as raters in the study, 29 subjects submitted complete diaries, recording their instructors' performance over a four week period. These 29 subjects also rated their instructors' performance at the end of this four week period.

Hypothesis 3 predicted a significant overall relationship between critical incidents and instructor ratings. The correlation between these ratings and the critical incidents for each of the dimensions were as follows: Organizational Skills, $\underline{r}(27)=.85$, $\underline{p} < .001$; Subject Relevance, $\underline{r}(27)=.48$, $\underline{p} < .01$; Student-teacher Relations, $\underline{r}(27)=.25$, $\underline{p} < .30$; and Communication Skills, $\underline{r}(27)=.85$, $\underline{p} < .001$. The overall correlation (i.e., collapsing across dimensions) between the ratings and critical incidents was found to be $\underline{r}(27)=.44$, $\underline{p} < .001$. Only four of the five original dimensions of instructors' performance were used here, due to the great overlap between critical incidents submitted for two of the dimensions (as well as the small proportion of incidents submitted for one of those two dimensions). That is, critical incidents submitted for the Subject Relevance dimension were very similar (and often identical) to those critical incidents submitted for the dimension of Knowledge of Subject Matter. Subjects frequently failed to submit incidents for the Knowledge of Subject Matter dimension. This may have been due to the perceived similarity between this dimension and the Subject Relevance dimension.

Comparing the "Validity" and "Reliability" Methods

Hypothesis 4 predicted a significant correlation in the rank ordering of rater accuracy estimates (i.e., the correlations between ratings and critical incidents scaled by CIRs) derived by the "validity" and the "reliability" methods. The demonstration of comparable results from the two methods would provide support for the use of one method as a "surrogate" measure for the other. This has very important practical implications, given that constraints of the real world would very likely prohibit the use of multiple observers of a given ratee's performance (as required by the "validity" method, but not by the "reliability" method). Therefore, it would be beneficial for those attempting to implement performance rating systems in organizations if it could be demonstrated that a method which requires only one observer ("reliability") yielded results comparable to a method requiring multiple observers ("validity").

The comparison of the two methods was made using the managerial data. In order to compare the two methods, for each of the four dimensions each of the accuracy scores derived for each rater by the "validity" method was correlated with each of the accuracy scores for each rater derived by the "reliability" method. Thus, two correlational matrices (one depicting correlations derived for each rater by the reliability method on each of the four dimensions and the other depicting correlations derived for each rater by the validity method on each of the four dimensions) were correlated with one another. This analysis of the two correlational profiles of raters yielded the following overall correlations (i.e., collapsed across raters) for each of the four dimensions: Structuring and Controlling the Interview, $\underline{r}(36)=.15$, $\underline{p}<.36$; Establishing and Maintaining Rapport, $\underline{r}(36)=.39$, $\underline{p}<.02$; Reacting to Stress, $\underline{r}(36)=.10$, $\underline{p}<.99$; and Obtaining Information, $\underline{r}(36)=.22$, $\underline{p}<.19$.

To further investigate Hypothesis 4, an analysis of variance was also performed using as the dependent variable the difference scores between: the "reliability" critical incidents and ratings (Method 1); the "validity" critical incidents and ratings (Method 2); and the manager true scores and ratings (Method 3). As can be seen in Table 2, this analysis revealed a significant Method effect, F (2,74)=9.48, $\underline{p}<.001$, indicating significant differences among these three methods. Multiple comparison tests revealed that Method 1 and 2 were not significantly different from each other, but were different from Method 3, lending some support to the generalizability of the reliability and validity methods. The analysis of variance also revealed a significant effect of Dimension, F(3,111)=13.38, $\underline{p}<.001$, as the previous correlational analyses would suggest. A post hoc multiple comparison test showed that the "Establishing and Maintaining Rapport" and "Obtaining Information" dimensions were not significantly different. These were significantly different from the other two dimensions, "Structuring and Controlling the Interview" and "Reacting to Stress." As can be seen in

Table 2

Analysis of Variance Using Difference Scores

Source	df	SS	MS	F	p
Between Subjects					
Rater	37	70.86	1.92		
Within Subjects					
Method	2	11.00	5.50	9.48	.001
Method X Rater	74	42.95	.58		
Dimension	3	45.39	15.39	13.38	.001
Dimension X Rater	111	127.82	1.15		
Method X Dimension	6	23.27	3.88	7.46	.001
Method X Dimension X Rater	222	115.81	.52		
Total	455	437.11			

Table 2, the analysis of variance also showed a significant Method X Dimension interaction. When a post hoc multiple comparison test of this interaction was performed, there were found to be a number of significant effects. Nonetheless, further inspection of these data failed to yield any discernible patterns. Because the error term for this interaction had such a large number of degrees of freedom, this was not surprising.

Thus, there appears to be .only very limited support from the correlational analyses of the generalizability of the methods, but the analysis of variance yields stronger support to this hypothesized relationship. However, the use of difference scores in an analysis of variance procedure has been criticized (e.g., Johns, 1981). The major problems, as Johns notes, lie in the unreliability of the difference scores and possible correlation of difference scores with their components and other variables. Johns also questions the meaningfulness of difference scores, given the way many researchers have used them (e.g., without regard for direction of the differences in scores). The points raised by Johns have been raised by other authors as well and, therefore, results of the previously reported analysis of variance in this study which uses difference scores must be interpreted with caution.

Given the lack of substantial support for Hypothesis 4, an investigation into the notion that the quality of critical incidents submitted by the 38 participants may have affected these results was begun. A sampling of critical incidents from the raters on all dimensions was rated for quality on a five-point scale by a group of seven Subject Matter Experts trained in the critical incident methodology. Subject Matter Experts were graduate students in Psychology and practicing Industrial Psychologists.

A *t*-test was performed to compare incidents submitted by the seven most accurate raters to those submitted by the seven least accurate raters and this analysis revealed no significant differences between the two groups ($t(55)=1.16$, $p < .59$). Also, the critical incidents written for the Establishing and Maintaining Rapport dimension (i.e., the dimension with the significant correlation between the two methods) showed no higher quality ratings than did critical incidents for any of the other dimensions.

Table 3 summarizes the results found for the four different methods of estimating rater accuracy. This table depicts the standard deviations and mean correlations between critical incidents and ratings (by dimension) for the managerial data (both the reliability and validity methods) and the instructor data, as well as the standard deviations and mean correlations between managerial ratings and true scores.

Assessing the Effects of Cognitive Complexity on Rater Accuracy

The effects of the cognitive complexity variable were assessed by correlating each individual rater's ratings (across managers) with the true scores for these managers provided by Borman (1977). Several students failed to complete the cognitive complexity scale, therefore only 26 subjects were used in this part of the study. A mean accuracy score (the correlation across managers) was derived for each rater across dimensions. These accuracy scores were then correlated (across subjects) with subjects' cognitive complexity scores. This analysis produced a nonsignificant $r(24)=.04$, $p < .99$, failing to support Hypothesis 5.

Hypothesis 5a predicted that accuracy estimates based on the instructor data would be significantly positively correlated with rater accuracy. To investigate this hypothesis, the same procedures as described immediately

Table 3

Mean Correlations Between Critical Incidents and Ratings By Dimension

Method	Dimension*	SD**	Mean
Reliability	1	.52	.68
	2	.55	.78
	3	.42	.71
	4	.63	.50
Validity	1	.51	.78
	2	.38	.76
	3	.46	.81
	4	.39	.69
Instructor	1	1.45	.85
	2	1.15	.48
	3	1.73	.25
	4	2.10	.85
True Score (correlation between true score and ratings)	1	.44	.72
	2	.38	.64
	3	.42	.81
	4	.43	.73

*For Reliability, Validity and True Score methods: Dimension 1 = Structuring and Controlling the Interview; Dimension 2 = Establishing and Maintaining Rapport; Dimension 3 = Reacting to Stress; Dimension 4 = Obtaining Information. For Instructor method: Dimension 1 = Organizational Skills; Dimension 2 = Subject Relevance; Dimension 3 = Student-teacher Relations; Dimension 4 = Communication Skills.

**Standard deviations are based on z scores; means are based on the correlation between critical incidents and ratings by dimension (averaged across subjects).

above were followed for the instructors' ratings, except that the accuracy estimates used to compute mean accuracy scores for each rater were those accuracy score correlations derived by the "reliability" method described previously (i.e., correlations between that rater's critical incidents and ratings). Again, the correlation between accuracy correlations and cognitive complexity scores ($\underline{r}(24)=.02$, $\underline{p} < .99$) was nonsignificant, failing to lend support to Hypothesis 5a.

Evidence for a Generalized Rating Ability

To determine if evidence of a generalized rating ability existed, rater accuracy estimates (as derived by correlating manager ratings with Borman (1977) true scores) were correlated with accuracy estimates based on these subjects' instructor data (derived by the "reliability" method). This correlation, a Spearman rho, was performed across raters, with $\underline{r}(27)=.13$, $\underline{p} < .50$. The correlation failed to support Hypothesis 6.

Assessing the Effects of an Individual Differences Variable on Generalized Rating Ability

The two sets of accuracy estimates described immediately above were converted to Fisher z scores, and averaged for each rater. Thus each rater had one mean accuracy score, representing his/her "generalized rating ability." Subjects' (raters') accuracy scores were then correlated with their cognitive complexity scores (across subjects), yielding an $\underline{r}(27)=.11$, $\underline{p} < .59$, failing to support Hypothesis 7.

Discussion

A number of hypotheses were not supported in the present study. The first hypothesis, which was the validation of the critical incident methodology as a procedure for determining the accuracy of raters, was supported, however. The significant correlation between the mean critical incident values and the managerial true scores provided support for the validity of the critical incident method as a suitable replacement for true scores. This was a particularly significant finding, since in the real world we have no "true scores" of individual performance, and thus must rely on other criteria for the purpose of validating performance ratings. This study suggests that the scaled critical incident technique is a viable alternative for this purpose. However, it was also hoped that individual accuracy estimates derived for raters based on these critical incidents could be generalized from a laboratory setting (where true scores were available) to a "real world" setting where no such scores were available. Unfortunately, Hypothesis 6, which predicted this generalized accuracy was not supported in the present study.

Borman's (1977) research used the same managerial videotapes to explore the notion of a generalized rating ability. Borman compared "within-task" consistency for accuracy to "across-task" consistency for accuracy. "Within-task" consistency was the consistency in accuracy scores for individual raters across dimensions and raters <u>within</u> a particular rating situation (e.g., videotaped managers' performance). To examine "across-task" consistency, Borman developed another series of videotapes depicting performers in a different job (i.e., job recruiters interviewing applicants). Across-task consistency was then defined as consistency in individual raters' accuracy scores across the two types of jobs. Borman found evidence of within-task consistency in rating accuracy (with correlations of accuracy scores for the recruiter and manager jobs being _r_ = .60 and _r_ = .65, respectively), but much less evidence of across-task consistency (_r_ = .46). The present study attempted to look at this across-situation consistency in rater accuracy by looking at two very different jobs in different settings and time frames. Accuracy scores derived in the controlled laboratory setting using seven minute managerial videotapes were correlated with those scores derived in a classroom situation occurring over a four week period. As noted previously, it was hoped that across-task consistency in accuracy could be demonstrated in a setting such as the classroom situation, certainly more similar to the kinds of situations individuals rating performance in the real world face. Unfortunately, the accuracy scores were not consistent across situations, with the correlation of the two sets of accuracy scores being _r_ = .13, $p < .50$.

Borman (1977) may have been correct when he suggested that "individual differences on 'abilities' associated with rating performance accurately may be situation specific" (p. 250). However, in the present study, other factors may have contributed to this failure to find consistency across rating tasks. Some of these will now be discussed.

One possible reason for the inconsistency in accuracy scores derived from the managers' videotaped performance and scores derived from the instructors' performance is that the dimensions used in the manager tapes may have been more behaviorally-oriented than the instructor dimensions. In considering the dimensions that were used for purposes of these two exercises, the argument might be made that the dimensions used in the videotaped exercises could be linked more easily to behaviors for which they were developed than could the instructor dimensions. For example, it may be easier to identify specific behaviors representing "Obtaining Information" (e.g., the manager asked the employee specific questions about particular events such as a fight with another employee) than those indicative of Subject Relevance. The latter dimension requires the rater to decide whether the subject matter discussed was germane to the objectives of the course. Nonetheless, the previously reported findings concerning quality of critical incidents indicate this was not the case. Subject Matter Experts rated instructor and manager critical incidents. The "quality" criterion used in these ratings concerned the specificity of behaviors reported (i.e., a "good" incident was one that reported details of an event that occurred, including only behaviors observed, whereas a "poor" incident was one that was vague and evaluative in nature). Mean quality ratings for the critical incidents written about managers were as follows: Structuring and Controlling the Interview, 2.351; Establishing and Maintaining Rapport, 2.357; Reacting to Stress, 2.801; and Obtaining Information, 2.506. Mean dimensional ratings for the critical incidents written about instructors were as follows: Organizational Skills, 2.959; Subject Relevance, 3.097; Student-teacher Relations, 3.474; and Communication Skills, 3.316. Thus, as can be seen from these findings, the critical incidents written about instructors were consistently rated as higher in quality than those for the managers. It would appear, therefore, that raters in the present study did not have greater difficulty in identifying specific behaviors when reporting the instructors' performance than the managers' performance.

It might be the case that the performance of the managers in the manager videotapes provided subjects with much more salient cues of performance indicative of the dimensions than did instructors observed by the subjects. Thus, even though instructor critical incidents were more specific and of higher quality, many may have been irrelevant to dimensions for which they were supposedly written. Limited anecdotal evidence of this possibility was provided in the fact that many subjects recorded the same incidents about managers (in almost identical words---such as quoting what the manager had said). This never happened with the instructor incidents, even though many students took classes together and observed the same instructors (although it must be noted that they had opportunity to observe a much larger range of performance).

Along this same line, it is very likely that variability in individual ratees' performance accounted for the discrepancy in the accuracy estimates. That is, subjects observed the managers only six to seven minutes, and there was little variation in each of the individual managers' performance during these episodes. Each of the individual instructors, on the other hand, was observed for a total of twelve hours, during which time it is very likely that more variability in performance was observed (e.g., an instructor who is usually very well organized might have had personal problems one day that kept him/her from adequately preparing for a particular class). The differences in opportunity to observe variation in individual ratees' performance in these situations may have affected the accuracy estimates. One way to determine if this is the case is to compare the variance in the ratings of the managers versus the instructors. Unfortunately, this cannot be done in this study since the identity of the instructors rated was anonymous. One area of research that deserves further attention is the effect of variation in ratee performance on performance ratings given. It is very probable that raters are affected quite differently by such variation, given what is known of individual differences in memory.

Another possible reason that support was not found for the notion of a generalized rating ability is that in the real world situation these subjects may have been more "ego-involved," -- that is, their personal feelings about the course or the instructor may have been tied to the incidents written or ratings made, whereas this may not have occurred with the Borman videotapes. While Borman's (1977) research was conducted in a controlled laboratory setting, with subjects not being acquainted with the "managers" (actors) on the videotapes, subjects rating instructors in the present study were face-to-face with the ratees they were evaluating, and in many cases probably already had personal feelings about the ratees as teachers. For example, over the four weeks, a subject may have recorded positive incidents about an instructor. This same subject may have taken a test in that instructor's class at the end of the four week period and received a low grade. Consequently, that subject may have given the instructor very poor ratings (reflecting the student's anger or disappointment and not the instructor's performance). This may have resulted in a low correlation for this subject on the instructor data, whereas the subject may have been very conscientious and received a high accuracy estimate on the managerial data. This might also happen, for example, if a subject felt that the Borman videotape exercise was just another experiment and was not conscientious about making ratings or recording incidents, but felt that the instructor exercise might in some way modify a situation. As an example, one subject in the study who complained that she did not like the Borman tape exercise was more enthusiastic about recording instructor critical incidents and asked to use the incidents to give feedback to a particular instructor. Thus, many such situations might account for the differences found in these two situations.

Regardless of the failure to demonstrate a generalized rating ability, given that the critical incident method was demonstrated to be a valid one, it was then possible to use this method to identify "valid" raters. When ratings were correlated with other observers' critical incidents, (i.e., the "validity" method), there was found to be a significant relationship between them, as predicted by Hypothesis 2.

This method of comparing other observers' critical incidents to ratings can be thought of as a measure of convergent validity, as previously discussed. Such validity estimates can be derived for individual raters within an organization. The demonstration of the validity of individual raters of performance is becoming increasingly important, in light of litigation involving performance appraisal systems over the past few years (e.g., Cascio & Bernardin, 1981).

As noted previously, however, the "validity" method requires the use of multiple raters, which is not likely to be feasible in most organizations. This study attempted to show that a method requiring only one observer of performance would suitably replace a method requiring more than one observer. While the critical incidents and ratings generated by the "reliability" method were significantly related (as predicted by Hypothesis 3), the data failed to demonstrate substantial evidence of a relationship between the "reliability" and "validity" methods, contrary to Hypothesis 4. That is, individual raters identified as "valid" by one method were not necessarily identified as "valid" using the other method when the same ratings were used (i.e., the managerial ratings). For example, Rater #1 had validity (accuracy) estimates of .43, .21, .35, and .01 for the four dimensions with the reliability method, but had accuracy estimates of .89, .85, .21, and .67 with the validity method. Appendix E depicts the "reliability" and "validity" accuracy estimates for individual raters (including those based on the manager data and those based on the instructor data) as well as accuracy estimates based on correlations of ratings with the managerial true scores. There were also definite differences in the generalizability of the reliability and validity methods among the four dimensions studied. As correlational analyses showed for one dimension, Establishing and Maintaining Rapport, accuracy estimates derived by the two methods were significantly related, whereas those for the other three dimensions were not.

The failure to find a significant relationship between the two methods may indicate that there is some systematic bias affecting the correlation in the reliability method. That is, it may be the case that at least some of the subjects were assigning ratings to the managers very much in line with the critical incidents they recorded for those managers (thus the high correlation), but, at the same time, perhaps were "cueing-in" only to certain aspects of the managers' performance. For example, a rater may have developed an initial unfavorable impression of a particular manager, and recorded only unfavorable incidents about that manager's behavior, ignoring positive aspects of that manager's behavior. The notion of existing impressions biasing final evaluations fits well with the Newell and Simon (1972) information processing model discussed in Appendix A. Briefly, this model proposes that raters of performance may be conceptualized as information processing systems which are required to solve a problem (i.e., make a performance rating) in a given task environment (i.e., the environment in which the ratee's behavior has been observed). Raters must process all incoming information concerning the ratee in the context of an already existing internal "network" of information consisting of previously received information about the ratee and rating situation, and the rater's own preferences, motivation and knowledge of rating. During this processing phase, new information is compared to previously existing information in this "network," after which it will either be stored intact along with existing information, reinterpreted to conform to previous information, or disregarded. Applying this model to the rating situation previously mentioned concerning the rater observing a particular manager, it is easy to see how biases such as initial unfavorable impressions can result in inaccurate ratings. This may occur even when the behavior has clearly exemplified a particular level of performance. That is, a rater may form an initial unfavorable impression of the manager, such that all subsequent information regarding the manager's superior performance is "reinterpreted" (distorted) to conform to this initial unfavorable impression that has been formed. If such information processing activities could actually be traced (as discussed in Appendix A), raters particularly susceptible to this "distortion method" of incorporating new performance information might be eliminated from any attempts to validate individual raters using this reliability method (i.e., a method dependent on one person's observations). It may even be possible to train individuals to become cognizant of such distortions and help eliminate them. Much more research is needed, however, to map the complex processes involved in arriving at a performance evaluation.

Nonetheless, examination of the original incidents indicates that the type of processing distortion described above appears to have occurred with a number of raters. Influenced by initial impression, the subject may have given the manager a very unfavorable rating. Thus, when the unfavorable incidents are scaled and correlated with the ratings, they will show a high positive relationship, yet not reflect the manager's true performance. When the "validity" technique is used, however, the number of observers of that particular manager is large enough that the effect of a number of such biased raters is negligible. The question then becomes, "What is the critical number of observers needed to render valid estimates of individual rater accuracy?" This is one area that needs to be explored in future research. It may be that "valid" individual raters could be initially identified by the use of the "validity" method with a small number of "other" observers. Once identified the organization could then rely extensively on these raters to appraise employees' performance (i.e., organizations might employ individuals who do nothing but rate others' performance). The validity of these raters might be periodically assessed by the "reliability" method as discussed above.

Neither Hypothesis 5 nor Hypothesis 5a (concerning the effect of cognitive complexity on rater accuracy) was supported in the present study. Hypothesis 5 predicted that cognitive complexity would be significantly related to the accuracy estimates derived for the managerial data, while Hypothesis 5a predicted that cognitive complexity would be significantly related to accuracy estimates derived for the instructor data. Correlational analyses used to test Hypothesis 5 and Hypothesis 5a failed to render support for these hypotheses. As noted in the introduction, cognitive complexity is an individual differences variable which has in the past been shown to affect ratings (or various qualities of ratings, such as halo error). One study has shown that raters high in cognitive complexity demonstrate less halo error than raters low in cognitive complexity (Schneier, 1977). It was expected that subjects identified as high in cognitive complexity in the present study would rate more accurately than subjects low in cognitive complexity. The point might be raised that perhaps the influence of cognitive complexity was not found in the present study (whereas it was in <u>one</u> earlier study) because the dependent variable used in this study (accuracy) is something very different from that used in Schneier's study (halo). However, halo error was also examined in the present study as in Schneier's work, and no effects were found for cognitive complexity. Thus, even when these data were examined using the dependent variable used by Schneier as well as an additional variable, no effects were found for cognitive complexity. Nonetheless, it should be noted that accuracy (as measured in the present study) depends on the viewpoint of a number of individuals--Subject Matter Experts, who are used to derive the true scores for the manager data and Critical Incident Raters, who are used to derive accuracy estimates for the instructor data, whereas halo measures are based solely on individual raters' assessments. While it may be reasonable to assume that a particular rater's accuracy is affected by his/her cognitive complexity (if accuracy could be measured by correlating that rater's ratings with the ABSOLUTE TRUTH), it may be the case that the involvement of other raters (SMEs and CIRs) masks this relationship. Given that there is typically a wide range of cognitive complexity scores within any SME or CIR group, it may be that the input of these subjects into the final accuracy measures influences the findings when accuracy is used as a dependent variable. More research is needed in which accuracy is used as a dependent variable to examine the effects of cognitive complexity on rating behavior.

One way that the effect of cognitive complexity on rater accuracy can be studied is to use written vignettes of hypothetical ratees' performance on very objective, quantifiable dimensions with true scores defined merely by "counting" ratees' behaviors, so that neither SMEs nor CIRs are needed. For example, a dimension such as absenteeism might be used (along with other such quantifiable dimensions) with the rater reviewing a large number of vignettes of a particular ratee's performance. The rater might be asked to rate the ratee, having to recall behaviors reported or represented by the vignettes. This rating could then be compared to the hard number data derived from the vignettes themselves. In this way, no input is required from other observers/raters of the ratee's performance. It may be that to look at the true effect of cognitive complexity on rater accuracy, only the rater in question each time can be used to derive accuracy measures. Studies such as the one suggested above provide one alternative for executing this type of research.

Hypothesis 7 predicted that a generalized rating ability (as operationalized by averaging accuracy estimates) would be significantly affected by rater cognitive complexity. This hypothesis was not supported. Again, the same argument applies here as was predicted above for failure to support Hypothesis 5.

Summary and Conclusions

The present study attempted to explore the use of a technique for validating individual raters of performance. Given that the demonstrated validity of ratings and individuals making those ratings has become of critical concern, the critical incident method discussed herein is of particular interest because, unlike many other validation strategies, it is a method whose implementation would be feasible for most organizations. The method was used to derive estimates of rater accuracy. Until the present time, most of the research done using rater accuracy estimates relied on "true scores" of performance, which are not available in real world situations (e.g., Borman, 1977). One of the purposes of the present study was to demonstrate that the critical incident method could be used as a suitable replacement for true scores in deriving accuracy estimates. This was shown to be the case, thus "validating" the critical incident method as a useful technique for identifying accurate raters within an organization.

Critical incidents were used to study rater accuracy in two different ways: First, they were used in what was termed a "validation" strategy (because the same performance was assessed by different methods--raters).

They were also used in a strategy that attempted to assess what might typically be termed "reliability" (the same performance was assessed by similar methods--raters). Critical incidents were shown to be significantly related to ratings made with each strategy. However, contrary to what was predicted, overall rater accuracy estimates derived by both methods were not shown to be significantly related. That is, accuracy estimates derived by the "reliability" method (a much more desirable method in terms of practical utility for organizations) were not found overall to be comparable to those derived by the "validity" method (a method in line with what is typically defined as construct validity, and more likely defensible in court). These two sets of accuracy estimates were, however, found to be comparable on at least one dimension of performance and further analyses suggested that the diemsnion in question may play a major part in the determination of the generalizability of the two methods.

The present study also examined the notion of a generalized rating ability by comparing accuracy estimates derived in two types of settings: a controlled laboratory setting where only a brief segment of performance is seen and a real world situation in which raters had an opportunity to observe performance for a much longer period of time. The notion of a generalized rating ability was not supported by the data, and possible reasons (such as differences in the types of dimensions and variability in ratee performance) were discussed.

The effects of cognitive complexity on both rater accuracy and a generalized rating ability were also examined. No support was found for the effect of cognitive complexity on either of these variables. As previonsly noted, it may be the case that in order to study individual rater accuracy, all "others" must be removed from the accuracy measure.

The results of the present study highlight the fact that there is still much research needed in the area of individual differences in rater behavior, especially those related to the processing of performance information in rendering a performance evaluation. The present study demonstrated a technique which can be used to identify the "accurate" raters within an organization. Given that some raters are better than others, what makes them better? Do they differ in their ability to perceive particular behaviors reflective of different levels of performance for different dimensions? Do they have different schema for processing new information concerning a ratee's performance? Do the differences in these raters lie in their ability to recall performance information? For example, are they differentially susceptible to primacy and recency effects? Are the raters affected differently by factors such as variability in ratee performance? Questions such as these raised by the present study point to the need to further investigate more process variables of performance appraisal (as opposed to the study of the end-products -- ratings and their psychometric characteristics). As mentioned previously, the need for increased attention in this area has been voiced by a number of researchers. Unless questions such as these are addressed in the literature, it is very unlikely that much progress will be made in increasing our understanding of the "end products" of appraisal which have previously received much attention in the literature. More study of these kinds of variables should greatly enhance our understanding of the entire evaluation process, and thus hopefully enable practitioners to implement more effective, valid performance appraisal systems which are coordinated with other personnel functions.

References

Banks, C. G. A laboratory study of the decision-making processes in performance evaluation. Unpublished dissertation, University of Minnesota, 1979.

Bass, B. M. Reducing leniency in merit ratings. Personnel Psychology, 1956, 9, 359-369.

Bass, B. M. & Turner, J. N. Ethnic groups' differences in relationship among criteria of job performance. Journal of Applied Psychology, 1973, 57, 101-109.

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

Bell, F. O., Hoff, A. L., & Hoyt, K. B. A comparison of three approaches to criterion measurement. Journal of Applied Psychology, 1963, 47, 116-119.

Bernardin, H. J. Implications of the Uniform Guidelines on Employee Selection Procedures for the Performance Appraisal of Police Officers. Proceedings of the National Workshop on the Selection of Law Enforcement Officers, C. D. Spielberger (Ed.), 1979, 97-102.

Bernardin, H. J., Orban, J. A., & Carlyle, J. J. The contribution of trust in the appraisal process and rater characteristics to performance ratings. Proceedings of the Academy of Management, 1981.

Bernardin, H. J. & Pence, E. C. Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 1980, 65, 60-66.

Bernardin, H. J. & Walter, C. S. Effects of rater training and diary keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Bigoness, N. J. Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. Journal of Applied Psychology, 1976, 61, 80-84.

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. Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 1977, 20, 238-252.

Borman, W. C. Individual differences correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 1979, 3, 103-115.

Borman, W. C., Hough, L. M., & Dunnette, M. D. Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error. ARI Technical Report, TR-78-A12, 1978.

Bronfenbrenner, U., Harding, J., & Gallwey, M. The measurement of skill in social perception. In D. C. McClelland, A. L. Baldwin, U. Bronfenbrenner, & F. L. Strodbeck, Talent and Society. Princeton: Van Nostrand, 1958, 29-111.

Bruner, J. S., & Tagiuri, R. The perception of people. In G. Lindzey (Ed.), Handbook of Social Psychology, Vol. 2, Cambridge, Massachusetts: Addison Wesley, 1954.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 75, 34-49.

Campbell, J. P. Psychometric theory. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Cascio, W. F. Applied psychology in personnel management. Reston, Virginia: Reston Publishing Co., 1978.

Cascio, W. F. & Bernardin, H. J. Implications of performance appraisal litigation for personnel decisions. Personnel Psychology, 1981, 34, 211-226.

Cascio, W. F. & Valenzi, E. R. Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 1977, 62, 278-282.

Chapman, L. J. & Chapman, J. P. Genesis of popular but erroneous psychodiagnostic observations. Journal of Abnormal Psychology, 1967, 72, 193-204.

Charest, A. G., Cowart, D. G. & Goodman, P. S. Multi-instrument, multirater, multitrait method for assessing measures of managerial performance. Experimental Publication System, 1969, 3, 092A.

Cline, V. B. Interpersonal perception. In B. A. Meher (Ed.) Progress in experimental personality research, Volume I. New York: Academic Press, Inc., 1964.

Clowers, M. R. & Fraser, R. F. Employment interview literature-- Perspective for the counselor. Vocational Guidance Quarterly, 1977, 26, 13-26.

Cooper, W. H. Ubiquitous halo: Implicit construct correlation theory versus BARS. Paper presented at the Meeting of the American Psychological Association, Montreal, Canada, 1980.

Cooper, W. H. Conceptual similarity as a source of illusory halo in job performance ratings. Journal of Applied Psychology, 1981, 66, 302-307.

Cronbach, L. J. Processes affecting scores on understanding of others and assuming "similarity." Psychological Bulletin, 1955, 52, 177-193.

Crow, W. J. & Hammond, K. R. The generality of accuracy and response in interpersonal perception. Journal of Abnormal and Social Psychology, 1957, 54, 384-390.

DeCotiis, T. L. & Petit, A. The performance appraisal process: A model and some testable propositions. Academy of Management Review, 1978, (July), 635-646.

DeJung, J. E. & Kaplan, H. Some differential effects of race of rater and combat attitude. Journal of Applied Psychology, 1962, 46, 370-374.

Dickinson, T. L. & Tice, T. E. A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 1973, 9, 421-428.

Elmore, P. B. & LaPointe, K. Effects of teacher sex and student sex on the evaluation of college instructors. Journal of Educational Psychology, 1975, 67, 368-374.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Ethnic group membership as a moderator of the prediction of job performance. Personnel Psychology, 1971, 24, 609-636.

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

Gaylord, R. H., Russell, E., Johnson, C., & Severin, D. The relations of rating to production records: An empirical study. Personnel Psychology, 1951, 4, 363-371.

Gellerman, S. W. The management of human resources. Hinsdale, Illinois: Drydan Press, 1976.

Goodman, P., Furcon, J., Ross, J. Examination of some measures of creative ability by the multitrait-multimethod matrix. Journal of Applied Psychology, 1969, 53, 210-213.

Gordon, M. E. The effect of the correctness of the behavior observed on the accuracy of ratings. Organizational Behavior and Human Performance, 1970, 5, 366-377.

Greenhaus, J. H. & Gavin, J. F. The relationship between expectancies and job behavior for white and black employees. Personnel Psychology, 1972, 25, 449-455.

Grey, R. J. & Kipnis, D. Untangling the performance appraisal dilema: The influence of perceived organizational context on evaluation processes. Journal of Applied Psychology, 1976, 61, 329-335.

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

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, N. J. Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 1974, 59, 705-711.

Hastorf, A. H., Schneider, D. J., & Polefka, J. Person perception. Reading, Mass.: Addison Wesley, 1970.

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

Ivancevich, J. M. A multitrait-multirater analysis of a behaviorally anchored rating scale for sales personnel. Applied Psychological Measurement, 1977, 1, 523-531.

Jacobs, R., Kafrey, D., & Zedeck, S. Expectations of behaviorally anchored rating scales. Personnel Psychology, 1980, 33, 595-604.

Johns, Gary. Difference score measures of organizational behavior variables: A critique. Organizational Behavior and Human Performance, 1981, 27, 443-463.

Jones, E. E., Shaver, K. G., Rock, L., Goethals, G. R., & Ward, L. M. Patterns of performance and ability attribution-- An unexpected primacy effect. Journal of Personality and Social Psychology, 1968, 10, 317.

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

Kane, J. S. Alternative approaches to the control of systematic error in appraisals. Paper presented at the Scientist-Practitioner Conference in Industrial/Organizational Psychology, Virginia Beach, Virginia, 1980.

Kane, J. S. & Lawler, E. E. Performance appraisal effectiveness: Its assessment and determinants. In B. M. Staw (Ed.), Research in Organizational Behavior, Volume I. Greenwich, Conn.: JAI Press, 1979.

Kavanagh, M. J. The content issue in performance appraisal: A review. Personnel Psychology, 1971, 24, 653-668.

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

Kelly, G. A. The psychology of personal constructs. New York: W. W. Norton, 1955.

Kenny, D. A. & Berman, J. S. Statistical approaches to the correction of correlational bias. Psychological Bulletin, 1980, 88 (2), 288-295.

Kipnis, D. Some determinants of supervisory esteem. Personnel Psychology, 1960, 13, 377-391.

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

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

Landy, F. J. & Farr, J. L. Performance rating. Psychological Bulletin, 1980, 87, 72-107.

Landy, F. J. & Trumbo, D. A. The psychology of work behavior. Homewood, Ill.: Dorsey Press, 1979.

Latham, G. P. & Wexley, K. N. Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 1977, 30, 255-268.

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

Lewis, N. A. & Taylor, J. A. Anxiety and extreme response preferences. Educational and Psychological Measurement, 1955, 15, 111-116.

London, M. & Poplawski, J. R. Effects of information on stereotype development in performance appraisal and interview context. Journal of Applied Psychology, 1976, 61, 199-205.

Mandell, M. M. Supervisory characteristics and ratings: A summary of recent research. Personnel Psychology, 1956, 32, 435-440.

McCormick E. J. & Tiffin, J. Industrial psychology. Englewood Cliffs: Prentice-Hall, Inc., 1974.

Miner, J. B. Explication of a method for finely graduated estimates of abilities. Journal of Applied Psychology, 1917, 1, 123-133.

Mullins, C. J. & Force, R. C. Rater accuracy as a generalized ability. Journal of Applied Psychology, 1962, 46, 191-193.

Meyers, L. S. & Boldrick, D. Memory for meaningful connected discourse. Journal of Experimental Psychology: Human Learning and Memory, 1975, 1 (5), 584-591.

Newcomb, T. M. An experiment designed to test the validity of a rating technique. Journal of Educational Psychology, 1931, 22, 279-289.

Newell, A. & Simon, H. A. Human problem solving. Englewood Cliffs: Prentice-Hall, Inc., 1972.

Nieva, V. F. & Gutek, B. A. Sex effects on evaluation. Academy of Management Review, 1980, 5 (2), 267-276.

Ronan, W. W. & Latham, G. P. The reliability and validity of the critical incident technique: A closer look. Studies in Personnel Psychology, 1974, 6, 53-64.

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

Schneider, D. J. Implicit personality theory. Psychological Bulletin, 1973, 79, 294-309.

Schneider, D. E. and Bayroff, A. G. The relationship between rater characteristics and validity of ratings. Journal of Applied Psychology, 1953, 37, 278-380.

Schneier, C. E. Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 1977, 62, 541-548.

Schneier, C. E. & Beatty, R. W. Influence of role prescriptions on performance appraisal processes. Academy of Management Journal, 1978, 21 (1), 129-135.

Scott, W. E., Jr., and Hamner, W. C. The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. Organizational Behavior and Human Performance, 1975, 14, 360-370.

Seashore, S. E., Indik, B. P. & Georgopoulos, B. B. Relationships among criteria of job performance. Journal of Applied Psychology, 1960, 44, 195-202.

Sechrest, L. & Jackson, D. N. Social intelligence and accuracy of interpersonal predictions. Journal of Personality, 1961, 29, 167-182.

Severin, D. The predictability of various kinds of criteria. Personnel Psychology, 1952, 5, 93-104.

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

Sherif, M., White, B. V., & Harvey, O. J. Status in experimentally produced groups. American Journal of Sociology, 1955, 60, 370-379.

Smith, P. C. & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-1555.

Taft, R. The ability to judge people. Psychological Bulletin, 1955, 52, 1-23.

Taft, R. The ability to judge people. In W. W. Ronan & E. P. Prien (Eds.), Perspectives on the Measurement of Human Performance. New York: Appleton-Century-Crofts, 1971.

Taguiri, R. Person perception. In G. Lindzey and E. Aronson (Eds.) Handbook of social psychology, Volume III. Reading, Mass.: Addison-Wesley, 1969.

Taylor, E. K., Parker, J. W., Martens, L. & Ford, G. L. Supervisory Climate and performance ratings: An exploratory study. Personnel Psychology, 1959, 12, 453-468.

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

Thorndike, R. L. Personnel selection. New York: Wiley Press, 1949.

Tucker, M. F., Cline, V. B. & Schmidt, J. R. Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. Journal of Applied Psychology, 1967, 51, 131-138.

Vannoy, J. S. Generality of cognitive complexity--simplicity as a personality construct. Journal of Personality and Social Psychology, 1965, 2, 385-3961

Waters, L. K. & Waters, C. W. Peer nominations as predictors of short-term sales performance. Journal of Applied Psychology, 1970, 54, 42-44.

Werts, C. E., Joreskog, K. G., & Linn, R. L. Analyzing ratings with correlated intrajudge measurement errors. *Educational and Psychological Measurement*, 1976, *36*, 319-328.

Wherry, R. J. *The control of bias in ratings: VII. A theory of rating*. Final Report No. 922, Army Research Institute for the Behavioral and Social Sciences, 1952.

Whitla , D. K. & Tirrell, J. E. The validity of ratings of several levels of supervisors. *Personnel Psychology*, 1954, *6*, 461-466.

Zedeck, S. & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 1972, *7*, 447-456.

Appendix A

Review of Literature Concerning the Rating Process and Variables Affecting Validity and Psychometric Characteristics of Ratings

The Rating Process

In recent years there seems to have developed a growing discontent with the disproportionate amount of attention given to rating products (e.g., resultant ratings, various rating formats) as opposed to the process of making a performance evaluation. In fact, several authors have called for less concern with products and more concern with process (e.g., Cooper, 1981). The need clearly exists for conceptualizing the rating process in some theoretical model. Cognitive and information processing approaches seem very feasible alternatives, especially if one considers the amount of information concerning the ratee's performance that must be observed, processed, stored, and later, retrieved by the rater in order to make a performance rating. Research in the memory literature seems especially applicable to performance appraisal situations, since the accuracy of the rating made depends so much on how well the rater is able to recall the ratee's performance over the course of the rating period. For example, there is evidence in the literature that serial position effects occur (especially recency effects) when the material to be remembered is complex (e.g., Meyers & Boldrick, 1975). Since performance observations may be considered fairly complex stimuli, one might expect to find primacy and recency effects in recall of such information when the rating is being made (i.e., the resultant rating would reflect a disproportionate weighting on performance observed at the beginning and the end of the rating period, since these observations are the ones remembered best by the rater). There is literature in the performance appraisal area to suggest that primacy effects do seem to influence performance ratings. For example, Jones, Rock, Shaver, Goethals, and Ward (1968) found that ratees whose performance was depicted in descending order tended to be rated as more able than ratees whose performance was depicted in ascending or random order (thus, performance occurring at the beginning of the appraisal period seemed to be remembered best). Also, Sherif, White, and Harvey (1955) and Mackie, Wilson, and Buckner (1954) have shown that perceptions (ratings) of performance become distorted in order to "fit" initial impressions concerning performance. Research on selection interviews adds support to this notion. The literature in this area has demonstrated that initial impressions about others can affect the manner in which subsequent performance of others is viewed. It has even been suggested that typically in a selection interview, the interviewer forms an impression of the interviewee in the first five minutes, and spends the remainder of the interview "searching" for information to corroborate this initial impression (Clowers & Fraser, 1977). Though these studies suggest the critical need for examining the amount of variance and ordering patterns of performance, there is unfortunately a paucity of research concerning the relative effects of performance distributions on resultant ratings. It has been suggested by a number of authors that differences in performance distributions affect the ratings of performance (e.g., Gaylord, Russell, Johnson, & Severin, 1951; Kane & Lawler, 1979). Nonetheless, very few studies have been done which have attempted to specifically ascertain the effects of differing performance distributions on ratings. One study which did attempt to do this found that _high variability_ in individuals' performance levels resulted in higher ratings on _ability_ to perform the task, but lower ratings on _motivation_ to perform the task, and when performance is presented in ascending order of effectiveness levels, ratees are rated as more motivated than those ratees whose performance is presented in descending or random orders of effectiveness (Scott & Hamner, 1975). Thus, there is some limited evidence that the order in which the rater views the ratee's performance (e.g., the levels of effectiveness seen initially, or at the end of the appraisal period) may affect resultant ratings. One of the purposes of the present study is to examine performance ratings for evidence of order effects, more specifically, recency effects. As mentioned previously, however, phenomena such as primacy or recency effects in ratings can be understood better if the rating process as a whole is conceptualized in some theoretical model. By conceptualizing the rating process in this way, it is possible to examine still other individual differences variables of raters which may affect resultant ratings, in addition to those associated with memory. Another purpose of the present study is to examine some of these individual differences variables in the context of a "generalized rating ability." The notion of a generalized rating ability will be discussed later. At the present time, however, it is important to first discuss a model of human information processing which may be applicable to the performance rating situation.

A Model of Rater Evaluation

A theoretical model especially relevant to the rating process is that offered by Newell and Simon (1972). Newell and Simon have proposed a fairly complex information processing system (IPS) model. Briefly, Newell and Simon have conceptualized human beings as information processing systems which receive information from the environment, operate on (process) this information and make final decisions/evaluations concerning this information. The emphasis of this model, however, is on the processing stage -- i.e., what processes are being utilized by the individual to reach the final evaluation, given particular information about the problem at hand. Even though outside observers cannot know how the individual represents the information received and the knowledge relevant to the task the individual already has stored, the model proposed by Newell and Simon enables one to conceptualize the sequence of events, so to speak, in arriving at a final evaluation. By doing this, it is hoped that further empirical study will enable one to identify actual processes being utilized by the individual. The theory of information processing proposed by Newell and Simon seems especially applicable to the performance evaluation task. As the model is presented below, specific references will be made to its application to performance appraisal.

Newell and Simon have postulated two basic parts of the problem-solving situation: the information processing system and the task environment. In a performance evaluation situation, the problem to be solved would be the final rating given to an individual by the rater; the information processing system would be the individual rater; and the task environment would be the environment in which the rater is required to appraise the performance of the ratee. The information processing system contains, then, four major components: _receptors_ (structures through which information is received); a _processor_ (the unit which performs various operations on the information being received); _symbol structures_ (sets of symbols connected to one another which enable one to represent new information in terms of previously established symbols); and _effectors_ (which initiate the appropriate action once the information has been processed). According to this conceptualization, what occurs in a performance rating situation is as follows: The rater observes the performance of the ratee and enters this new information into the information processing system. This information is then translated into an internal representation of the task environment (i.e., the rater constructs the rating situation internally via performance information observed). Appropriate symbols and symbol structures are selected, as well as the appropriate operators (processes) (i.e., the rater, recognizing the evaluation that must be made, decides how this new information will be operated on in light of already existing information relevant to this particular ratee, rating situation, and the rater's own preferences, motivations, and knowledge of rating). At this point, the rater may compare new information to old information about the ratee; decide to disregard this new information; reinterpret this new information to make it "fit" with previously stored information; or store this new information along with previous information. Disregarding or reinterpreting this new information seem likely possibilities for the rater to choose, especially if the rater has already formed an impression of the ratee. As mentioned previously, there is evidence in the literature that suggests observers (raters) of performance form initial impressions and distort (or reinterpret) subsequent observations to conform to these initial impressions.

The next step in this evaluation process is to select a problem-solving method (i.e., the rater decides how to combine all this information to reach a final evaluation) and then apply it to the existing knowledge state. If the solution results in the rater's goal (a final rating), the rater has finished. If not, however, the rater may do one of three things: 1) apply a different problem-solving method to the knowledge state; 2) generate a totally new internal representation of the task environment (i.e., look for new information concerning the ratee's performance and/or rating situation in order to "see the picture in a new light," so to speak); 3) abandon the problem-solving task (rating task) altogether.

As noted previously, the purpose of conceptualizing the performance rating process in this way is to attempt to identify the actual program (i.e., operators, problem-solving methods) being used by the rater in making the final evaluation. Attempts to identify these programs have traditionally involved process tracing procedures. These procedures involve methods that attempt to "track" the individual as he/she is actually processing the information. Methods employed in this endeavor have included such techniques as recording verbalizations made by the rater during the time when considering the performance observations while attempting to make a rating. Supposedly, this technique enables one to pinpoint what information (observations) were most attended to and how information was combined in making the final judgment. Banks (1979) used such an experimental procedure when she attempted to determine the information attended to and the processes used by raters in reaching final evaluations concerning performance. In her study, she attempted to capture the decision-making processes of 156 business students who rated videotaped performance of ratees. In order to trace raters' decision-making processes, Banks looked at such variables as frequency of certain judgments (ratings), range of ratings made, latency of rating decisions, and raters' ability to make ratings which discriminated among the various ratees. Banks had raters view videotapes of ratees' performance (see Borman, Hough, & Dunnette, 1976) and indicate the point at which they were making judgments while watching the videotapes by pressing buttons (corresponding to anchors on rating scales) and verbally recording their reasoning for making such judgments. Results indicated that raters generally attended to different aspects of performance when making judgments. It was also found that raters who attended to some of the _same_ aspects of performance rated these aspects differently (even though anchored rating scales were used). Thus, this study suggests that differences _do_ exist among raters in the way they process observed performance information.

Another technique that has been used to identify procedures being used by raters to process performance information is policy capturing. Policy capturing is a regression analysis technique which determines the relative weights assigned to different pieces of information (e.g., dimensions, observations) by the rater in making an overall rating.

By conceptualizing the individual's rating process in such a way, it is easier to consider different variables which may enter into this process and consequently affect the appropriateness of the final decisions made (i.e., appropriateness in terms of the validity of the final decisions). This conceptualization is one way of structuring the empirical information that currently exists concerning variables which affect the validity of ratings, as these variables apply to the individual rater. By viewing these data in this way, these variables can be made more meaningful in their relevance to the problem of identifying valid raters. The literature concerning these variables which have been shown to affect performance ratings will now be reviewed. The variables discussed will deal with different facets of the information processing system model, falling into three major categories: characteristics of the rater (e.g., abilities, motivation, implicit theories concerning performance); characteristics of the ratee (e.g., abilities and performance levels); and situational and organizational variables (i.e., those related to the task environment such as amount of trust in the organization in which the rating is being done, position of the rater in relation to the ratee).

Variables Affecting Ratings

One assumption that must be made when using ratings of performance (as opposed to purely objective indices of performance, for example) is that the raters who are observing others' performance and evaluating this performance are capable of making rational, unbiased decisions based on those observations. However, it has been suggested that objectivity in performance judgments is questionable because "raters subscribe to their own set of assumptions" (Cascio, 1978, p. 320). In fact, it has even been suggested that performance ratings may reflect the "implicit personality theories" of the raters, rather than the observation of actual behaviors of the ratee (Landy & Farr, 1980). Implicit personality theory, which may be seen as one aspect of the person perception phenomenon, suggests that raters tend to form ideas about how certain dimensions of behavior (or attributes) are related and apply these hypotheses when making ratings (Newcomb, 1931; Schneider, 1973). Therefore, raters may be giving ratings based on assumptions of their own, rather than actual observations of behavior. This would at least partially explain the halo effect sometimes found in ratings, i.e., the correlation between behaviors or traits which the investigator believes to be independent (Cooper, 1980; Kenny & Berman, 1980). There is evidence in the literature that this phenomenon exists (e.g., Bruner & Tagiuri, 1954; Chapman & Chapman, 1967; Schneider, 1975).

In addition to these assumptions that raters may hold about the ratees, raters also differ on other characteristics which may adversely affect the psychometric qualities and accuracy of the ratings given. Demographic characteristics of raters which have been found to affect ratings have included such factors as rater sex (e.g., Elmore & LaPointe, 1974; London & Poplawski, 1976; Nieva & Gutek, 1980); race of the rater (e.g., DeJung & Kaplan, 1962); and age of the rater (Klores, 1966; Mandell, 1956). Results are mixed as to whether rater experience has an effect on ratings given to subordinates. Jurgenson (1950), Mandell (1956) and Gordon (1970) found that rater experience significantly affected ratings given, but Cascio and Valenzi (1977) and Klores(1966) found no such differences among ratings given. Performance level of the rater has also been associated with the quality of ratings given (e.g., Bayroff, Haggerty, & Rundquist, 1954; Schneider & Bayroff, 1953). The leadership and management styles of raters have been reported to have significant effects on ratings given as well (Klores, 1966; Taylor, Parker, Martens, & Ford, 1959).

Several personality and mental traits of raters have been studied to determine their effect on raters in making unbiased judgments. For example, Mandell (1956) found that raters who had low self-confidence tended to give less lenient ratings than those high in self-confidence. In another study, raters who displayed high levels of anxiety were found to use extreme response categories of rating scales instead of using the middle categories (Lewis & Taylor, 1955). Borman (1974) correlated a number of personality measures with rating accuracy and found that accurate raters tended to be relatively high in self-competence, detail-oriented, high in tolerance, empathetic, non-aggressive, and high in intelligence (including verbal reasoning ability and high grades). Although Borman's work suggests intelligence/aptitude may be one of the best predictors of rating accuracy, perhaps even stronger results might be shown if other measures were used (e.g., measures previously documented in behavioral research such as SAT scores or GPA).

The cognitive complexity of the rater was also found by Borman to be related to rater accuracy, though not as strongly as in other studies (e.g., Schneider, 1977). Perhaps one reason for these somewhat less compelling findings is that Borman used a measure of cognitive complexity not typically used in such studies, i.e., one derived from the Kelly Reperatory Grid (Kelly, 1955). Other measures of cognitive complexity might demonstrate more positive findings.

Another important characteristic of the rater which may affect the accuracy or psychometric characteristics of the ratings given which deserves mention is the motivation of the rater to provide an accurate rating. Several authors have noted that the rater's motivation to make an accurate judgment can be affected by such factors as: consequences of the ratings given, appropriateness/adequacy of the rating scale used, current standards of performance (i.e., their availability and appropriateness), and purpose of the appraisal (DeCotiis & Petit, 1978; Taft, 1971). Kane (1980) has distinguished between two types of rating error based on this important motivational factor. According to Kane, motivated errors, or errors due to "deliberate misrepresentations" of what the rater has observed and is aware of concerning the performance of the employee may be alleviated by allowing raters to participate in the development of the performance appraisal system. However, it should be noted that this may result in scales of poor quality. As several authors have shown, idiosyncratic perceptions of the job for which the performance appraisal system is being developed may result if an unrepresentative sample of employees help develop the system (Bernardin, 1979; Borman, 1974; Schneier & Beatty, 1978). Therefore, a precaution that might be taken to minimize rating errors is to ensure a representative sample of employees was involved in the development of the system.

Other methods recommended by Kane (1980) for controlling the misrepresentation of actual performance included modifying the outcomes for good versus bad ratings, or disguising the values of the ratings given (e.g., using an instrument such as the forced-choice scale). Kane argues that nonmotivational errors, or "unintended inaccuracies" are affected by many of the rater characteristics discussed previously, and can be alleviated somewhat by training. However, as Bernardin and Pence (1980) showed, training raters to eliminate the most commonly studied errors such as leniency and halo may render the ratings inaccurate.

Nonetheless, as has been shown in the above discussion, factors which produce these psychometric errors, as well as those contributing to inaccuracy in ratings are important and must be taken into consideration when ratings are used to measure performance.

The second major category of variables associated with rating errors is that of characteristics of the ratee. As with rater characteristics, certain demographic factors have been significantly related to ratings given, such as ratee sex (e.g., Nieva & Gutek, 1980) and race of the ratee (Bass & Turner, 1973; Farr, O'Leary & Bartlett, 1971; Greenhaus & Gavin, 1972). Probably the most important ratee variable, in terms of determining the ratings given, is the actual performance exhibited by the ratee (and rightfully so!). However, various aspects of a ratee's performance may account for some error found in ratings. For example, although several authors cite performance level as accounting for the largest proportion of variance in ratings given (e.g., Bigoness, 1976; Gordon, 1970), certain aspects of that performance such as variability in performance level have been shown to significantly affect ratings given. For instance, as noted previously, Scott and Hamner (1975) found that variability in individuals' performance levels resulted in higher ratings on ability to perform the task, but lower ratings on motivation to perform the task.

Cascio (1978) has also suggested that interactions between rater and ratee chracteristics may contribute to the error in ratings. For example, a few studies have shown that raters tend to give significantly higher ratings to ratees of their own race (e.g., Bigoness, 1976; Hamner, Kim, Baird, & Bigoness, 1974). Also, women tend to rate other women less favorably than men unless the performance criteria used in ratings are clearly specified (Nieva & Gutek, 1980).

Finally, various situational and organizational factors have been found to adversely affect the accuracy and psychometric qualities of ratings. The position of the rater in relation to the ratee (e.g., supervisor, peer, subordinate) has been shown to significantly affect ratings given, probably because of the opportunity to observe the ratee's job behavior, and amount of job-related contact the rater is able to have with the ratee (e.g., Borman, 1974; Klimoski & London, 1969). Along this same line, the availability of appropriate performance standards has also been linked to resultant ratings (DeCotiis & Petit, 1978). Consequences of the appraisals (i.e., whether ratings will be used for research purposes, employee development, or as a basis for various personnel decisions) have also been shown to have a significant effect on ratings given (Gellerman, 1976; Sharon & Bartlett, 1969).

The extent to which raters trust the appraisal system in use has also been suggested to affect the leniency of ratings given (Bass, 1956). In a study which examined the effect of both individual and organizational variables on appraisal ratings, it was found that while cognitive complexity of the rater accounted for only nine percent of the variance in resultant ratings, trust in the appraisal process accounted for 32% of the variance (Bernardin, Orban, & Carlyle, 1981). Bass (1956) has suggested several reasons why trust might affect ratings in this way. One possibility for leniency in ratings where there is low trust is that rating an employee's performance as poor might reflect upon the supervisor's management abilities (and the supervisor is usually the rater). Also, a supervisor can gain rewards for his/her employees by giving good ratings, thus promoting good relations between him/her and the employees and possibly increasing control over these employees. Also, the rater may feel a need to "approve" of others (thus rating leniently) either to gain approval for himself/herself or because he/she feels society demands such a response. According to Bass, the rater may also be projecting feelings about his/her own performance onto the ratings given to subordinates. Lastly, the rater may be rationalizing that any employee who would actually deserve an unfavorable rating would already have been removed from the organization.

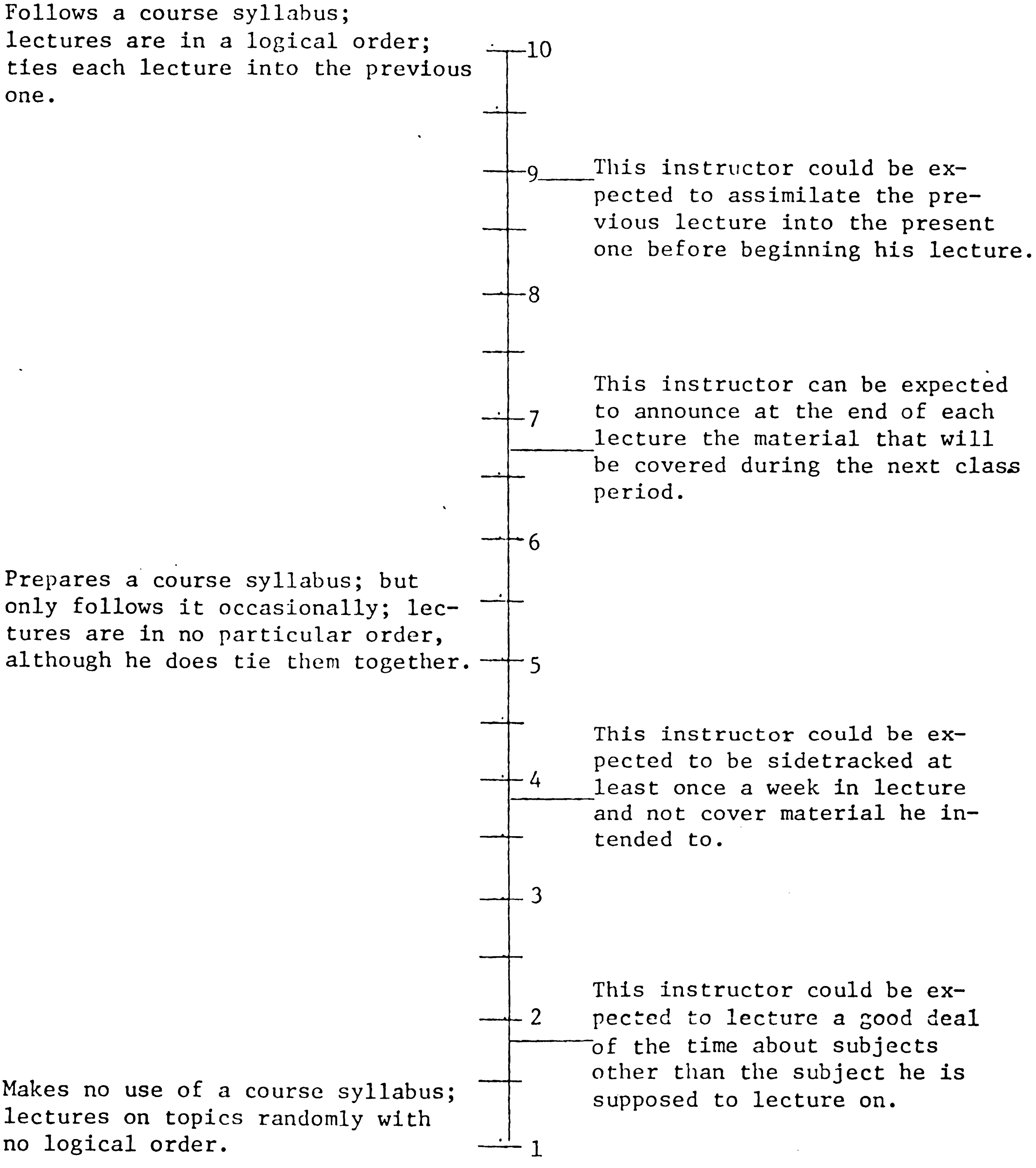
Certain situational factors have been linked to rater leniency as well, such as propinquity, social setting, and expressions of criticism (Kipnis, 1960), and presence of a non-compliant worker (Grey & Kipnis, 1976). Rating formats used have accounted for only 4-8% of the total variance in ratings, even though much of the research in performance appraisal has been devoted to the development and comparison of different rating formats (Landy & Farr, 1980). However, even though formats alone have been shown to account for little of the variance, formats may interact in some way with a generalized rating ability. This interaction may be an important aspect of the evaluation process to be studied in the future. Lastly, based on a review of research in performance appraisal, Wherry (1952) postulated that organizational and situational factors such as the variability in the work setting, machine-paced versus man-paced operations, union restrictions, complexity of the job, record keeping by the rater, and time lapse between observation of behavior and completion of the ratings all can serve as sources of error in ratings.

Thus, it is evident from the literature reviewed above that there are many potential sources of error in ratings. It has been noted that various aspects of the rating situation (e.g., rating format) which have received a great deal of attention in the literature do not account for nearly as much of the rating variance as do variables associated with the rater. Many different characteristics of the rater can influence the ratings given, and thus the variance the rater contributes to the ratings. A similar situation is found in the testing literature, i.e., individuals assessing responses to essay test questions contribute much of the variance to the final scores (Werts, Joreskog, & Linn, 1976). As pointed out, differences in raters' accuracy in assessing others' performance may not lie solely in abilities, but may be due to many other variables associated with individual raters.

Appendix B

Behaviorally Anchored Rating Scales for Rating Instructors

ORGANIZATIONAL SKILLS: A good constructional order of material; slides smoothly from one topic to another; design of course optimizes interest; students can easily follow his organizational strategy; course outline followed.



Follows a course syllabus; lectures are in a logical order; ties each lecture into the previous one.

9 — This instructor could be expected to assimilate the previous lecture into the present one before beginning his lecture.

7 — This instructor can be expected to announce at the end of each lecture the material that will be covered during the next class period.

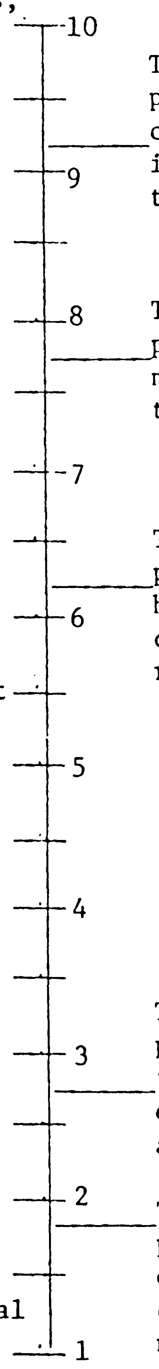
Prepares a course syllabus; but only follows it occasionally; lectures are in no particular order, although he does tie them together.

4 — This instructor could be expected to be sidetracked at least once a week in lecture and not cover material he intended to.

2 — This instructor could be expected to lecture a good deal of the time about subjects other than the subject he is supposed to lecture on.

Makes no use of a course syllabus; lectures on topics randomly with no logical order.

SUBJECT RELEVANCE: Relating of the subject matter to things important and meaningful to students; generalizes material to real world; distinguishes useful information from the trivial; applies subject matter to other areas of knowledge.



Relates course material to important facets of students' lives; lectures on material that is meaningful and useful without falling back on unimportant material

This instructor could be expected to take important principles in his subject area and illustrate them to his students through real-life experiences.

This instructor can be expected to discuss recent material which relate to the topics being discussed.

This instructor can be expected to relate the material he is teaching to another course which he knows is a requirement for his students.

Sometimes relates material to students; but only at their request lectures on material ranging from important to trivial.

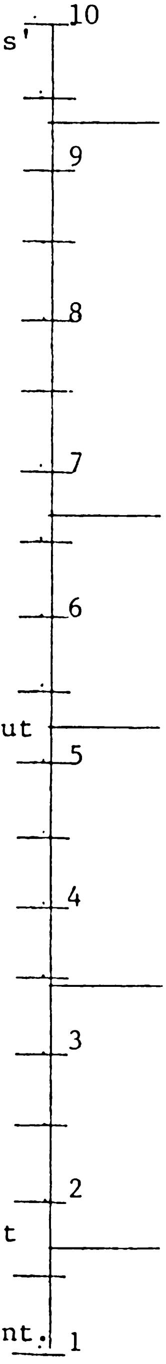
This instructor could be expected to pay little attention to local or national current events though they may be applicable to his course.

This instructor could be expected to spend a good deal of time ranting about political affairs and degrading national figures though the topic has no relevance to the class.

Doesn't distinguish trivial material from important; never relates course material to students and their interests, ideas, or problems.

STUDENT-TEACHER RELATIONS: An instructor's rapport with his students; respects the comments and suggestions of students; makes an effort to get to know his students on a more personable level; takes an interest in students as individuals; sensitive to students' needs and problems.

Has basic consideration for students; is interested in students' ideas and problems.



This instructor can be expected to make an extra effort to encourage students to ask questions.

This instructor can be expected to always recognize the students in his medium size classes although he sometimes forgets their names.

Makes no special attempt to know students personally; is attentive to students' ideas and problems, but only as they correspond to his ideas and expectations.

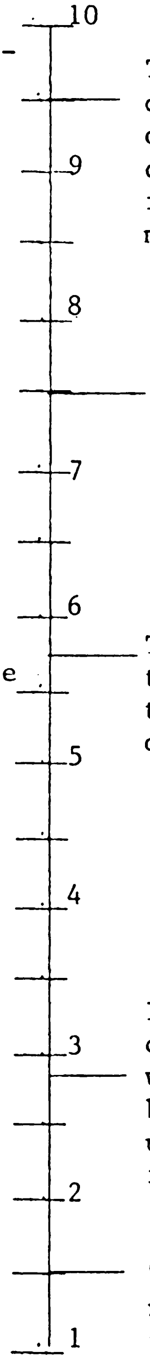
This instructor could be expected to tell students who have many questions to see him during his office hours.

This instructor could be expected to be emphatic about test dates and assignment due-dates regardless of any extenuating circumstances.

Sees students only as students, not as individuals, uninterested in students' ideas and problems, discourages any personal involvement.

This instructor could be expected to have a "superior" attitude in dealing with his students, making the student feel quite uncomfortable.

COMMUNICATION SKILLS: An instructor's ability and method of conveying material; delivery facilitates an easy understanding; no nervous habits that interfere with the learning process; a reasonable speech rate; good inflection in his voice.



Speaks clearly; is comfortable and at ease lecturing; is easily understood both lecturing and answering questions.

This instructor has a clear, distinct, excellent voice and can be heard anywhere in the classroom. He speaks with inflection and conveys each mood of the material.

This instructor can be expected to stop often during a lecture to repeat an idea so that it is clear.

Sometimes lectures are unclear and confused; at times is uncomfortable and distracting when lecturing.

This instructor can be expected to only occasionally talk too fast for the students to follow adequately.

In order to study for an exam of this instructor, students would rely much more on the book because they can't understand most of what the instructor says.

Is uncomfortable lecturing; has distracting nervous habits; has difficulty expressing lecture material clearly.

This instructor reads from his notes and speaks in a low monotone. It is almost impossible not to become drowsy during class.

Appendix C

Summated Scales for Rating Critical Incidents Written About Instructors and Managers

Rating Scale for Instructors

only slightly agree	agree a little	mildly agree	agree some-what	agree in part	tend to agree	moder-ately agree	gener-ally agree	agree on the whole	pretty much agree
1	2	3	4	5	6	7	8	9	10

DIMENSION A - ORGANIZATIONAL SKILLS

A good constructional order of material; slides smoothly from one topic to another; design of course optimizes interest; students can easily follow organizational strategy; course outline is followed.

DIMENSION B - SUBJECT RELEVANCE

Relating of the subject matter to things important and meaningful to students; generalizes material to the real world; distinguishes useful information from the trivial; applies subject matter to other areas of knowledge.

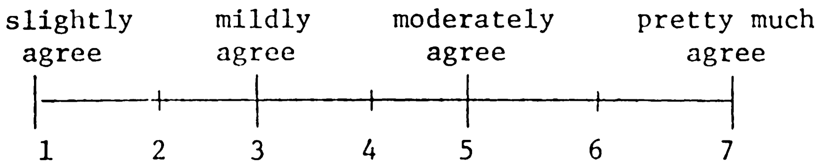
DIMENSION C - STUDENT-TEACHER RELATIONS

An instructor's rapport with the students; respects the comments and suggestions of students; makes an effort to get to know the students on a more personal level; takes an interest in students as individuals; sensitive to students' needs and problems.

DIMENSION D - COMMUNICATION SKILLS

An instructor's ability and method of conveying and delivering the material; delivery facilitates an easy understanding; no nervous habits that interfere with the learning process; a reasonable speech rate; good inflection in his/her voice.

RATING SCALE



DIMENSION A - Structuring and Controlling the Interview

The manager maintains control over the interview by ensuring that the interviewee doesn't control the interview when inappropriate, by clearly stating the purpose of the interview and being organized and prepared for the interview.

DIMENSION B - Establishing and Maintaining Rapport

The manager sets an appropriate climate for the interview (i.e., a nonhostile, nonbelligerent climate) by opening the interview in a warm, nonthreatening manner and being sensitive toward the interviewee.

DIMENSION C - Reacting to Stress

The manager reacted appropriately to stress during the interview by remaining calm and cool (even during the interviewee's outbursts), reacting reasonably to the interviewee's complaints, and appropriately sticking to his position when confronted by the interviewee.

DIMENSION D - Obtaining Information

The manager probed effectively into the interviewee's perceptions of problems so that meaningful topics were raised by asking appropriate questions and seeking solid information versus glossing over problems.

Appendix D

Diary for Recording Critical Incidents of Instructors' Performance

Organizational Skills: A good constructional order of material; slides smoothly from one topic to another; design of course optimizes interest; students can easily follow organizational strategy; course outline is followed.

Professor A - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor A - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Please list any other critical incidents you wish to add on the remainder and back of this page. (Please remember to identify the incidents according to professor.)

Subject relevance: Relating of the subject matter to things important and meaningful to students; generalizes material to real world; distinguishes useful information from the trivial; applies subject matter to other areas of knowledge.

Professor A - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor A - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Please list any other critical incidents you wish to add on the remainder and back of this page. (Please remember to identify the incidents according to professor.)

Student-Teacher Relations: An instructor's rapport with the students; respects the comments and suggestions of students; makes an effort to get to know the students on a more personal level; takes an interest in students as individuals; sensitive to students' needs and problems.

Professor A - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor A - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Please list any other critical incidents you wish to add on the remainder and back of this page. (Please remember to identify the incidents according to professor.)

Communication Skills: An instructor's ability and method of conveying and delivering the material; delivery facilitates an easy understanding; no nervous habits that interfere with the learning process; a reasonable speech rate; good inflection in his/her voice.

Professor A - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor A - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor B - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #1
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Professor C - Incident #2
What happened just before the incident?
What happened right after the incident?
What were the specific details of the incident?

Please list any other critical incidents you wish to add on the remainder and back of this page. (Please remember to identify the incidents according to professor.)

Appendix E

Accuracy Estimates for Individual Subjects

Appendix G

Accuracy Estimates for Individual Subjects

	Method 1 (Reliability)				Method 2 (Validity)				Method 3 (Instructor)				Method 4 (True Score & Rating)			
Subject	-Dimension*				Dimension				Dimension				Dimension			
Number	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	.43	.21	.35	-.01	.89	.85	.21	.67	.78	.31	.97	.99	.85	.70	.49	.89
2	.76	.71	.93	.76	.45	.67	.77	.25	.91	-.73	.94	.99	.85	.30	.61	.85
3	.81	.85	.61	.76	.94	.91	.89	.86	.99	.67	.68	.39	.68	.59	.91	.71
4	.64	.78	.08	.73	.72	.89	.96	.25	.72	.60	-.95	.98	.40	.92	.95	.16
5	.65	.85	.70	.38	.86	.62	.70	.88	.98	.27	.89	.86	.71	.59	.68	.81
6	.03	.72	.79	.64	.91	.58	.76	.85	.51	.51	-.99	-.99	.75	.30	.81	.80
7	.74	.60	.57	.76	.85	.88	.87	.72	.96	.66	.90	.97	.91	.58	.87	.57
8	-.02	.95	.90	.59	.51	.86	.88	.69	.95	.32	-.53	.94	.61	.51	.78	.84
9	.48	.92	.82	.77	.79	.81	.20	.75	.81	.96	.16	-.29	.79	.82	.20	.93
10	.47	.37	.74	.21	.93	.79	.64	.65	.31	-.56	-.82	.96	.78	.52	.72	.69
12	.01	.77	.92	.75	.90	.60	.87	.56	.99	.56	-.56	.94	.70	.53	.88	.34
13	.82	.56	.52	-.41	.80	.60	.97	.45	.30	.87	.95	.99	.89	.45	.91	.77

*For Method 1, Method 2 and Method 4: Dimension 1 = Structuring and Controlling the Interview; Dimension 2 = Establishing and Maintaining Rapport; Dimension 3 = Reacting to Stress; Dimension 4 = Obtaining Information. For Method 3: Dimension 1 = Organizational Skills; Dimension 2 = Subject Relevance; Dimension 3 = Student-teacher Relations; Dimension 4 = Communication Skills.

	Method 1				Method 2				Method 3				Method 4			
	(Reliability)				(Validity)				(Instructor)				(True Score & Rating)			
Subject	Dimension				Dimension				Dimension				Dimension			
Number	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
14	.70	.94	.90	.73	.39	.63	.55	.54	-.06	-.56	.97	.45	.16	.91	.73	.72
15	.87	.24	.44	.39	.91	.62	.95	.75	.55	-.07	.72	.87	.83	.38	.96	.55
18	.91	.85	.51	.86	.89	.76	.90	.88	.45	-.25	-.52	.15	.76	.63	.85	.80
19	.83	.04	.69	.37	.78	.39	.82	.29	-.71	.99	.46	.98	.51	.22	.91	.24
20	.37	.32	.72	.75	.75	.63	.83	.61	.59	.44	.96	-.17	.55	.45	.81	.81
21	.81	.90	.43	.64	.45	.80	.63	.63	.63	.45	.99	.99	.27	.93	.76	.79
22	-.64	.75	.84	.53	.37	.71	.81	.88	.99	-.03	-.91	-.99	.61	.75	.86	.96
23	.65	.58	.56	-.84	.31	.25	.58	.36	.81	.41	.75	.72	.20	-.01	.44	.52
24	.79	.97	.10	.07	.93	.76	.98	.56	-.81	.83	-.14	.90	.86	.77	.96	.56
27	.94	.84	.80	.59	.97	.91	.73	.71	.94	-.50	-.64	.67	.77	.84	.77	.55
29	.16	.78	.70	.08	.72	.85	.92	.77	-.26	-.27	-.99	.13	.44	.57	.97	.74
32	.42	.83	.85	.32	.50	.71	.84	.90	.98	.99	.00	.99	.49	.60	.79	.90
33	.73	.60	.53	.97	.91	.08	.76	.21	-.74	.29	.08	.99	.87	.43	.66	.49
34	.92	.96	.62	-.15	.89	.78	.68	.87	.99	.44	.95	.99	.82	.67	.66	.70
35	.74	.71	.75	.64	.88	.91	.85	.94	.99	.65	-.82	,67	.87	.58	.78	.94
36	.83	.27	.86	.68	.95	.69	.61	.77	.87	.29	-.51	.12	.95	.79	.59	.69

Subject Number	Method 1 (Reliability) Dimension 1	2	3	4	Method 2 (Validity) Dimension 1	2	3	4	Method 3 (Instructor) Dimension 1	2	3	4	Method 4 (True Score & Rating) Dimension 1	2	3	4
37	.79	.96	.91	.84	.09	.84	.50	.37	.72	-.59	.00	-.04	-.03	.64	.57	.54
11	.78	.97	.78	-.02	.39	.71	.83	.62	**	**	**	**	.49	.33	.77	.78
16	.55	.67	.74	-.01	.45	.93	.81	.64	**	**	**	**	.41	.73	.85	.61
17	.97	.74	.86	.63	.49	.85	.83	.60	**	**	**	**	.61	.66	.76	.53
25	.75	.92	.88	.40	.85	.93	.94	.67	**	**	**	**	.73	.66	.84	.51
26	.80	.82	.13	.93	.82	.87	.74	.92	**	**	**	**	.95	.61	.76	.95
28	.48	.88	.76	.35	.87	.90	.88	.76	**	**	**	**	.76	.62	.87	.67
30	.72	.58	.64	-.03	.90	.58	.58	.60	**	**	**	**	.77	.47	.53	.51
31	.79	.56	.75	.72	.69	.68	.93	.72	**	**	**	**	.81	.46	.80	.85
38	.67	.43	.20	-.15	.53	.62	.73	.33	**	**	**	**	.49	.91	.77	.33

**These subjects did not complete the second part of the study.

The three page vita has been removed from the scanned document. Page 1 of 3

The three page vita has been removed from the scanned document. Page 2 of 3

The three page vita has been removed from the scanned document. Page 3 of 3

An Investigation of a Method For Validating Individual Raters
Of Performance and Its Implications For a
Generalized Rating Ability

by

Jamie J. Carlyle

(ABSTRACT)

The present study explored the use of a technique for validating individual raters of performance and its implications for the existence of a generalized "ability" in raters to make accurate assessments of others performance. Subjects were asked to record critical incidents of ratees' performance in two types of job situations-- 1) a videotaped presentation of managers interviewing problem employees, and 2) instructors teaching in actual college classrooms. Subjects also rated the performance of these managers and instructors. Scaled critical incidents were correlated with ratings to derive three kinds of accuracy scores. Two sets of these accuracy scores (the managerial "reliability" and "validity" estimates) were compared to determine if a method for inferring validity using many raters' observations were comparable to a method using only one rater's observations. The accuracy scores derived in two types of settings (i.e., reliability estimates derived from manager data and reliability estimates derived from instructor data) were compared to determine the generalizability of rating accuracy across situations. Unfortunately, little empirical support was provided for the equivalence of the two methods (i.e., "reliability" and "validity") or for the generalized ability notion. Possible reasons for the failure of the present study to support the hypotheses are discussed, with emphasis on the

importance of considering the process of rating performance rather than the end products of such a process.