3-1

THE RELATIONSHIP BETWEEN COGNITIVE COMPLEXITY AND THE USE OF VARIOUS TYPES OF RATING SCALE FORMATS

by

MARY ANNE LAHEY

B.S., Illinois State University, 1976

A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1980

Approved by:

Major Professor

THIS BOOK
CONTAINS
NUMEROUS PAGES
WITH THE ORIGINAL
PRINTING BEING
SKEWED
DIFFERENTLY FROM
THE TOP OF THE
PAGE TO THE
BOTTOM.

THIS IS AS RECEIVED
FROM THE
CUSTOMER.

Acknowledgements

I would like to express my appreciation to the members of my thesis committee: Dr. Frank (Skip) Saal, Dr. Ronald Downey, and Dr. Richard Harris, for their help in the preparation and completion of this project. Special thanks go to both Skip and Ron for their strong support during the early stages of my initiation into the field of Industrial Psychology. Their continued encouragement has proven to be invaluable in the course of my professional, as well as personal, development.

TABLE OF CONTENTS

LIST OF TABLES

Research in the area of rating scale formats and rating behavior has yielded confusing, and often contradictory, results (Landy & Farr, 1980). Scale formats that have appeared, on an intuitive level, to have promise in the reduction of undesirable psychometric characteristics have been empirically disappointing. Behaviorally anchored rating scales (Smith & Kendall, 1963), mixed standard rating scales (Blanz & Ghisselli, 1972), and summated rating scales (Bernardin, Alvares & Cranny, 1976) all have their bases in the precise designation of the behaviors to be rated and are, thereby, supposedly less open to errors of interpretation. These formats, however, have not consistently emerged as more psychometrically sound than other types of rating scales (Bernardin et al, 1976; Finley, Osburn, Dubin, & Jeanneret, 1977; Saal, 1979; Schwab, Heneman, & DeCotiis, 1975). For this reason, research in recent years has tended to emphasize less the scale formats themselves and has focused to a greater extent on those aspects of the rating situation, or characteristics of the raters, that may influence the quality of ratings.

This trend is particularly noticeable in the context of work-related ratings, or performance appraisals. Many situational and individual difference variables have been suggested, or shown to be of importance, in affecting the process and results of assessing other persons' performance. For example, the raters' experience with a particular rating scale format (Borman & Vallon, 1974), the raters' participation in the development of the rating scales (Friedman & Cornelius, 1976), raters' training with and knowledge of the psychometric characteristics of ratings (Bernardin & Walter, 1977; Ivancevich, 1979), the organizational levels of the raters (Klimoski & London, 1974), the raters' own levels of performance (Kirchener & Reisberg, 1962), and the purpose

(experimental vs. administrative) of the ratings (Warmke & Billings, 1979) have all been found to be related to the psychometric quality of workers' performance ratings. In addition, early work (Estes, 1938; Kelly & Fiske, 1958) investigated the possibility that raters' level of intelligence (I.Q.) was related to their rating behavior. Although this work has been shown to be inconclusive in determining the relationship between I.Q. and ratings (Taft, 1962), the cognitive abilities of raters have not been abandoned in the search for potentially important aspects of rating behavior.

More recently, Schneier (1977) has suggested that the cognitive structure of raters may play an important role in predicting the psychometric properties of their ratings, as well as in moderating the raters' preferences for particular rating scale formats and their confidence in their judgements using different formats. Using relatively simple and complex rating scale formats, Schneier showed that the interaction between a rater's cognitive complexity and the complexity of the scale format may be an important factor in the attainment of psychometrically sound ratings. In light of the trends toward identifying relevant rater and situational characteristics as predictors of rating quality, these findings may prove to be extremely important. A more complete summary of these results will be presented following a general discussion of the construct of cognitive complexity, as well as relevant research.

Cognitive Complexity

The concept of cognitive complexity can be traced to George A. Kelly's (1955) theory of personal constructs. This theory is based on the notion that reality is not directly revealed to us and that it is subject to many different constructions. The ways in which reality can be construed are as

numerous as we, ourselves, can invent. In an effort to assimilate, interpret, and anticipate our environments, we each develop a system of constructs and impose them on the events we experience. This system of personal constructs provides cohesion or unity to the individual's world. Kelly contends that, although an event may be open to several different interpretations, some ways of construing it will be ultimately more useful than others for the anticipation of its recurrence (Adams-Webber, 1979). In addition, people must be understood in relation to their present attempts to anticipate environmental events--their personal construct system. The fundamental postulate of Kelly's (1955) theory is: "A person's processes are psychologically channelized by the ways in which he anticipates events" (p. 46).

This postulate and its eleven corollaries form an intricate theory of personality, and a more general theory of psychological processing. The basic unit of this theory is the personal construct. A construct is defined as "a way in which some things are construed as being alike and yet different from others" (Kelly, 1955, p. 105). Constructs are considered to be bipolar in nature. It is assumed that persons employ constructs that have constructual opposites, and that they are not restricted to the rules of formal logic which assert that an opposite can only be stated as the absence of the construct. For example, black vs. white may be a meaningful construct for conceptualizing the world, and there is no reason to demand (or even believe) that black vs. not-black and white vs. not-white are the more appropriate constructs to be used. Furthermore, it is assumed that each construct used by a person has a limited range of convenience, and that the construct is not useful or relevant outside of that range. The elements or events that fall into this range of convenience define the context of the construct. Similar constructs may have

different contexts or ranges of convenience for different persons; this represents another indication of the individualistic nature of the personal construct system.

It is important to keep in mind that the elements comprising the context of a particular construct do not constitute or define the construct. The construct is the fundamental unit within which elements or events are understood. The constructs represent the interpretations of situations and are, therefore, relatively abstract representations of thought processes. This level of abstraction is further exemplified by the contention that "it is not possible for one to express the whole of his construction system" (Kelly, 1955, p. 110). There are some constructs that are useful for conceptualization, and yet they defy any attempts to be meaningfully put into words. Similarly, it may be impossible for one person to fully understand the constructs used by someone else, since one's own constructual system may hinder the interpretation of another's system. These types of problems have led to one of the major focal points of personal construct theorists, that of systematically assessing individual differences in cognitive structures (personal construct systems).

Directly resulting from attempts to measure and analyze the personal construct systems of individuals has come the concept of, and the term, cognitive complexity. Cognitive complexity has generally been used to describe the ways in which individuals' thoughts are organized and integrated, as reflected by the system of constructs they use to describe and anticipate events in their environments. Bieri (1955) coined the term "cognitive complexity" and used it specifically to indicate the degree of differentiation that can be found within a person's construct system. For example, persons

who tend to perceive elements or events in an identical manner on several ostensibly different constructs are deemed cognitively simple or undifferentiated; whereas, persons who tend to perceive elements or events differently on every construct are deemed cognitively complex or differentiated (Adams-Webber, 1979). Cognitive complexity, therefore, is a concept that has mainly been intended to reflect the relative degrees of differentiation of persons' construct systems, and may be defined as the capacity to perceive the environment in a multidimensional manner (Bieri, Atkins, Briar, Leaman, Miller, & Tripodi, 1966).

One major problem with the research in the area of cognitive complexity derives from the fact that not all theorists, or measuring instruments, retain the concept of differentiation as the primary characteristic of cognitive structure. For example, Crockett (1965) used an operational definition of cognitive complexity that differed considerably from the measure of Bieri (1955), while maintaining the conceptual definition of differentiation. Bannister (1960), on the other hand, relied on the degree of integration among constructs, their interrelation with one another, as opposed to their relative differentiation, as his measure of cognitive complexity. Smith and Leach (1972) focused on a hierarchical analysis, an analysis of the finer details of the constructual system, for their conception of the construct. Other definitions of cognitive complexity have centered around the articulation of the constructs (the degree to which one can discriminate within a construct), the enumeration of the constructs (the total number that can be independently expressed), as well as various combinations of all these conceptualizations (Bieri et al, 1966). Because different measuring instruments were developed to reflect these different

conceptual definitions of cognitive complexity, the process of comparing and interpreting the research results derived from studies based on these different measures is difficult.

In spite of these difficulties, it is possible to point to particular areas of human behavior that appear to be moderated by the cognitive complexity of individuals (regardless of its precise definition and measurement). In his initial formulation, Bieri (1955) argued that persons whose construct systems reflect differentiated perceptions of the environment would have greater degrees of discrimination among their personal constructs and would, therefore, have a greater predictive sense of environmental events. He specifically proposed that his measure of cognitive complexity, the modified Role Construct Repertory (REP) Test (to be discussed in detail later), would correlate positively with accuracy in the prediction of others' behavior (Adams-Webber, 1979).

This contention was borne out in a number of subsequent studies of social judgement; more cognitively complex persons have been found to be more accurate in their predictions of others' behavior than less complex persons. This tendency has been found in experimental situations (Bieri, 1955, 1961) where subjects were asked to predict the responses of two acquaintances on social questionnaires. Leventhal (1957) found that cognitively complex subjects were slightly better at predicting others' behavior in social, work and school situations. He also found that cognitively complex persons perceived fewer similarities between themselves and others than did cognitively simple subjects, and that the complex subjects differentiated more among persons whose responses they attempted to predict. Plotnick (1961) found that cognitively complex social work students were better able to clinically assess their clients' difficulties, and were better at predicting the behavior

of those clients than were cognitively simple social work students. Other research (Adams-Webber, 1969, 1973) suggests that persons high on measures of cognitive complexity are better able to infer, from the behavior of others, their (the others') personal construct systems. Craig and Duck (1977) found that cognitively complex subjects were more accepting of others whose personal construct systems differed from their own.

Other kinds of abilities have also been found to be related to cognitive complexity. Greater degrees of cognitive complexity have been shown to be associated with a person's ability to discriminate among multidimensional stimuli (Tripodi & Bieri, 1964). This relationship was demonstrated when both the quality and the quantity of the information were varied. Campbell (1960) found that subjects who manifested low levels of cognitive complexity (cognitively simple subjects) were more likely to dichotomize elements and events, rather than to report their perceptions along dimensional continua. Olson and Partington (1977) reported results suggesting that cognitively complex subjects are better able to see events from the perspective of others, and are better able to "reconcile simultaneously the perceptions of oneself and others." Finally, cognitive complexity was shown to be related to the ability of subjects to integrate opposite (both positive and negative) information about target persons. Cohen (1961), Zajonc (1960) and Nidorf and Crockett (1965) all reported that cognitively complex persons were better able to accept potentially contradictory information in descriptions of other persons. Mayo and Crockett (1964) demonstrated that less complex persons showed a greater tendency toward both primacy and recency effects in impression formation, depending on the order and the valence of the information presented. Cognitively complex persons, on the other hand, reported impressions that were more ambivalent, being neither extremely positive nor

negatively biased, reflecting an integration of the contradictory information. Replications of this result have been reported by other researchers as well (Leventhal & Singer, 1964; Press, Crockett & Rosenkrantz, 1969).

A point worth repeating here is that these studies were often conducted using different measures of complexity. This may be regarded as both a potential strength and a potential weakness of the concept. Its strength may lie in the argument that cognitive complexity is a very useful framework for understanding a wide range of social judgement and social perception phenomena, and that this framework is not limited by the instruments used to measure it. The drawback lies in the argument that cognitive complexity is so broad a concept, and open to so many different interpretations, that it may be impossible to identify its parameters or to determine when it may be applicable (or inapplicable) to other research areas. In short, the construct validity of cognitive complexity may not be established. If this is, in fact, the case, the value of cognitive complexity as an explanatory tool is severely diminished. The possibility that cognitive complexity is tautological to other psychological concepts is addressed in a number of studies that have attempted to examine the relationship between it and other, more familiar, notions (e.g., intelligence, sociability, etc.). Although cognitive complexity is related to some of these, it (complexity) appears to remain a relatively distinct view of cognitive structure.

Initial studies investigating the relationship between cognitive complexity and other cognitive variables focused on intelligence. A number of studies correlated scores of academic intelligence/performance (SAT, ACE, verbal analogies) with measures of cognitive complexity and found no significant relationship between the two sets of scores (Crockett, 1965).

Thus, it would appear that cognitive complexity is a concept that is distinct from the general notion of academic intelligence. Cognitive complexity has also been compared with measures of "social intelligence," the effectiveness of persons' interpersonal relationships (Sechrest & Jackson, 1961). Although there were modest positive correlations among the measures of cognitive complexity and "social intelligence," these relationships were only reliable for two of the four complexity measures employed in this study. The degree to which cognitive complexity reflects a more general notion of interpersonal effectiveness, therefore, remains unclear.

In an investigation into the generality of cognitive complexity as a personality variable, Vannoy (1965) administered a battery of tests to individuals in order to appraise their level of cognitive complexity and other seemingly similar or related concepts. This study compared subjects' scores on 20 different tests, four of which were specifically designed to measure cognitive complexity. The remaining sixteen tests, designed to measure other cognitive skills, included measures of verbal and quantitative ability, category width, independence of judgements, social distance, intolerance of ambiguity and trait inconsistencies, assumed similarity among opposites, and an F-scale measure of authoritarianism. Correlations among the various measures were moderate to low, ranging from .45 to -.56. Correlations among the complexity measures ranged from .23 to -.43. A principal-axes factor analysis of these tests yielded six interpretable factors accounting for 91.4% of the common variance and 35.0% of the total variance. This procedure did not yield a large first factor on which all of the 20 tests, or even a substantial proportion of them, were substantially loaded, implying that cognitive complexity may not be a general personality

trait. Vannoy (1965, p. 385) concluded that cognitive complexity "may consist of a number of distinct, possibly independent, tendencies; not all of which are educed by any of the present measuring instruments."

Cognitive Complexity and Rating Behavior

As discussed above, cognitive complexity appears to be a concept that reflects relatively unique properties of an individual's cognitive structure. It is most often defined as the ability to discriminate among dimensions of stimuli, and is often expanded to encompass discriminations within those dimensions. Using these definitions, it follows that cognitive complexity involves very similar skills and abilities to those that are required for tasks involving the rating of others. In fact, the original measure of cognitive complexity, Kelly's (1955) Role Construct Repertory (REP) Test, is a complex rating task on which individuals are asked to rate significant others on their (the subjects') own personal constructs. In addition, the role that cognitive complexity appears to play in the processes of social judgement lends further support to the contention that cognitive complexity may play an important moderating role in rating behavior. Ratings in the context of performance appraisals constitute a large portion of the rating literature, and it seems reasonable to assume that cognitive complexity may be an important rater characteristic in these settings.

Schneier (1977) reported a study that examined the interaction between raters' cognitive complexity and their ratings of others' work performance using two different rating scale formats. Each rater's level of cognitive complexity was measured using a modified grid form of the REP test (Bieri et al, 1966). This measure requires the rating of ten individuals corresponding to role titles specified on the grid, using ten bipolar adjectives

also specified on the grid. The uniqueness of the ratings for each role-person on all constructs determines the individual's cognitive complexity. (A more detailed description of this procedure will be provided later.)

Raters were then required to rate the job performance of one, two, or three of their co-workers using each of two rating scale formats. One, a fourteen-dimension behaviorally anchored rating scale, was considered to be complex (cognitively demanding) in that it required raters to make fine discriminations among rating dimensions (each dimension was to be considered independently), and also required fine discriminations within each dimension (a 7-point linear scale was used). The other rating scale format, described as the "alternate scale," was considered to be relatively simple in the cognitive demands it placed upon the raters. The "alternate scale" contained only ten dimensions, thereby requiring less discrimination among dimensions. A 3-point (Above Average, Average, Below Average) ordinal scale was used, requiring less within-dimension discrimination than the behaviorally anchored rating scale.

High- and low-complexity rater groups were formed by splitting the complexity scores at the median, and *t*-tests between the groups were used to compare the ratings with respect to leniency, halo and restriction of range. Interactions between the raters' cognitive complexity and the complexity of the rating scale formats were found for leniency and restriction of range analyses. Cognitively complex raters showed significantly less leniency (mean ratings closer to the scale midpoint) when using the complex scale format; whereas, cognitively simple raters showed less leniency with the simpler scale format. Similarly, restriction of range (mean standard deviations of ratings across ratees) was reduced when the

raters' cognitive structures and the rating scale formats were compatible. Cognitively complex raters produced larger standard deviations (less range restriction) using the complex scale format and cognitively simple raters produced larger standard deviations using the simpler scale. In the case of halo error (standard deviations across rating dimensions), the cognitively complex raters showed less bias (greater deviations among ratings) than did the simple raters, irrespective of the rating scale format used.

In addition to these typical indices of rating quality, the study reported by Schneier (1977) investigated the "operational utility" of the two types of rating scales. Operational utility was measured by having raters report their levels of confidence in judgements using each scale and their satisfaction with the type of information about performance provided by each format. Operational utility was also measured by raters' preference for appraisal format in actual use. Two-factor analyses of variance (ANOVAs) were computed on the confidence and satisfaction scores yielding significant Complexity x Scale Format interactions. This indicates that the complex raters were more satisfied and confident using the complex scale and the simple raters were more satisfied and confident using the simpler rating scale. Contingency tables for rater preference showed a significant statistical association between format preference and rater complexity. Again, complex raters preferred the behaviorally anchored rating scale to the alternate scale, and the simple raters preferred the alternate scale to the behaviorally anchored scale format.

Although these findings may prove to be beneficial in explaining the contradictory results in the literature on behaviorally based rating scales, this study appears to be plagued by several problems that make it difficult to interpret the precise relationship between cognitive

complexity and rating behavior. First, the procedure used to measure cognitive complexity, analysis of the modified grid form of the REP test, bears a striking resemblance, both conceptually and operationally, to the definition of halo error typically found in the performance appraisal literature (Bieri, 1955; Borman, 1975). In order to calculate a person's complexity score, the ratings given to each role-person on the grid are compared across constructs. The greater the differentiation (variation) in these ratings, the greater is the person's cognitive complexity. That is, the more each construct is uniquely used to describe the ratee, the higher is the complexity level. This measure is very similar to the concept of halo. Halo is typically defined as a rater's inability or unwillingness to discriminate among dimensions of a given ratee's behavior (Saal, Downey, & Lahey, in press). It, therefore, seems reasonable that cognitively complex raters will, in fact, demonstrate less halo regardless of scale format, and will fail to show an interaction with rating scale format as Schneier (1977) hypothesized.

Second, the modified REP measure of cognitive complexity focuses solely on the concept of differentiation among personal constructs. Because cognitive complexity has been defined in a number of other ways, and has been expanded to encompass more than discriminating among constructs (e.g., the ability to discriminate within a single construct), it is unclear whether Schneier's (1977) results would be replicated using other measures of cognitive complexity. This complexity measure is, itself, a rating task, and it is not surprising to find that it is related to another form of rating behavior. Rating ability, however, is not the only aspect of cognitive complexity and if complexity is a general trait or concept, it should be reflected in other types of behaviors. This contention is

supported by the fact that studies (Schneier, 1979; Vannoy, 1965) show little or no relationship among different measures of complexity, and yet all of these measures had previously been shown to be related to social judgements. It may be that a measure of complexity that is less a function of ratings would facilitate identification of the specific skills or abilities that influence cognitive complexity, and thereby moderate rating behavior.

Third, the "alternate" rating scale format used in Schneier's (1977) study has not been particularly useful for the evaluation of work performance and has largely disappeared from the rating literature. Although a 3-point scale is easy to develop and use, it has been found to be error-prone in actual practice. More information needs to be obtained on the degree to which cognitive complexity interacts with other types of rating scales, particularly those more commonly found in the contemporary performance appraisal literature.

Another potential problem related to the rating scale formats utilized by Schneier (1977) revolves around the number of performance dimensions tapped by the scales. The behaviorally anchored rating scale and alternate rating scale differed in the number of dimensions tapped, fourteen and ten respectively. The rating analyses, particularly with respect to confidence, satisfaction and preference for a rating scale, are potentially confounded by this difference. It is impossible to tell if the interactions between complexity and rating scale format are a function of the overall complexity of the scales, as Schneier concludes, or solely a function of the number of dimensions tapped by each.

Finally, it should be noted that the analyses of the psychometric properties of the ratings were based on separate t-tests for each of the

rating scales. If the compatibility theory of rating behavior that Schneier (1977) proposes depends on complexity x scale format interactions, as the leniency and range restriction analyses suggest, suitable tests of these hypotheses need to be performed. Two-factor ANOVAs seem to be statistically more appropriate for the investigation of the proposed relationships between cognitive complexity and various types of rating scale formats.

Before the practical implications of this line of research can be discussed, it is important to note that Schneier's (1977) results are far from definitive. Studies that have attempted to replicate his findings with other samples of raters have usually failed to find any relationship between a rater's cognitive complexity and rating behavior (Schneier, Note 1). Bernardin and Boetcher (Note 2) also examined the relationship between raters' cognitive structure and ratings obtained using behavioral expectation scales. They found no significant differences between high- and low-complexity rater groups with respect to leniency and halo measures. Although the measure of complexity was the same as that used in the previous study, there were some differences in scale format and scale format development that may account for the differential results. The behavioral expectation scale used by Bernardin and Boetcher consisted of only seven dimensions, whereas Schneier's included fourteen dimensions. It may be that discriminations among dimensions do not become particularly difficult for cognitively simple raters until a certain critical number of dimensions are tapped. Also, Bernardin and Boetcher developed their scale in accordance with the procedures outlined by Bernardin, LaShells, Smith, and Alvares (1976), which have been shown to be useful in the reduction of rating errors. The quality of the behaviorally based rating scale may be a crucial variable

in determining the type of Complexity x Scale Format interactions that are reflected in the psychometric indices of rating quality.

These types of discrepancies are not unusual in the literature on rating scale formats (Landy & Farr, 1980). It, therefore, seems even more important to further examine the relationship between cognitive complexity and the specific properties of rating scale formats. It is particularly important to determine the kinds of variables that characterize a complex rater and distinguish him or her from a cognitively simple rater. It is similarly important to determine the variables that render a rating scale format complex, and distinguish it from a relatively simple rating scale. Without more information on these topics, attempts to apply Schneier's (1977) findings would be somewhat premature.

Purpose of the Present Study

The present study was designed to further examine the relationship between cognitive complexity and rating behavior. In particular, it attempted to identify the precise nature of the skills and abilities which comprise a complex, as opposed to a simple, rater by using a variety of techniques for measuring cognitive complexity that vary systematically along two dimensions. In addition, it was anticipated that the use of various rating scale formats, including those scales commonly found in current performance appraisal research, would provide insight into the specific aspects of rating scales that render them simple or complex.

In order to tap various aspects of raters' cognitive structures, three different measures of cognitive complexity were administered to each ratee. These measures included: 1) a modified grid form of the REP test introduced by Bieri et al. (1955); 2) a form of the original REP

test using a factor analytic scoring procedure suggested by Levy and Dugan (1956); and, 3) a sorting task measure of cognitive complexity proposed by Scott (1962). These particular measures were chosen because they seem to represent varied approaches to the measurement of cognitive complexity. Detailed descriptions of these measuring techniques are provided below.

The interactions of these cognitive complexity measures with different types of rating scale formats were investigated by asking raters to provide performance evaluations using each of four rating scale formats. The rating scales employed in this study included: 1) a behaviorally anchored rating scale (BARS); 2) a mixed standard rating scale (MSS); 3) a graphic rating scale (GRS); and, 4) a 3-point alternate scale (AS) modeled after the simple scale described by Schneier (1977). Except for the alternate scale, these scales represent those commonly found in the contemporary performance appraisal literature. It was anticipated that inclusion of these scales would lead to a clearer understanding of the specific aspects of rating scale format that interact with cognitive complexity to moderate rating behavior. The rating scales and procedures used in their development are discussed in detail below.

Ratings were analyzed, using analyses of variance, with respect to three typical indices of rating quality: leniency, the comparison of mean rating levels; halo, the dispersion of ratings across performance dimensions; and, restriction of range, the dispersion of ratings across ratees. The effects of primary interest were the main effect of Complexity (do high- and low-complexity raters differ with respect to each index?), and the Complexity x Scale Format interactions, (is cognitive complexity differentially related to different types of rating scale formats?).

METHOD

Instruments

Measures of Cognitive Complexity

The measures of cognitive complexity described here vary in two ways: the degree to which they depend upon rating behavior as an index of cognitive complexity and the origin of the personal constructs used in the assessment of cognitive complexity. Both of these dimensions are potential sources of information in determining the specific nature of the cognitive complexity-rating scale format relationship. Each of these measures has been found to be related to the accuracy of social judgements, although the relationships among the measures have not been firmly established.

Modified grid form of the REP test. This form of the Role Construct Repertory (REP) Test was designed for group administration by Bieri and his colleagues (1966). It consists of a 10 x 10 rating grid on which the columns represent persons being rated and the rows represent the rating constructs. The ten columns are identified by role labels (e.g., mother, friend, yourself) selected to be representative of meaningful persons in an individual's social environment. Beside each of these labels, the rater lists the names (or initials) of the persons in their life who best correspond to those roles. No name may appear on the grid more than once. The ten rows of the grid are prelabeled with sets of bipolar adjectives (e.g., outgoing-shy, independent-dependent) that serve as the basis for the ratings of the persons in each column. A 6-point rating scale is used, ranging from +3 to -3 (zero is excluded).

Cognitive complexity is measured by comparing each rating for a particular role-person with their ratings on all other constructs. In

comparing any two ratings, a point (1) is given if there is exact agreement (i.e., the same numerical value) on the two constructs. If the two ratings are not numerically equivalent, no points are given. This comparison is carried out for all possible rating combinations for each ratee, and the points are summed to give a total complexity score. Since there are 45 possible row combinations for each ratee, the highest score (indicating the least complexity) on a 10 x 10 grid is 450. A score of 450 would indicate that the rater was highly undifferentiated and gave the same numerical rating to each ratee on the ten constructs.

In order to reduce response set errors and increase the reliability of the grid scores, two modifications of this procedure suggested by Vannoy (1965) were employed. Adjectives describing the rating constructs were counterbalanced on the scale so that all positive or negative adjectives did not fall on the same side (left, right) of the scale, and the positive (+) and negative (-) signs preceding the numerical scores were replaced with R (for right-hand side of the scale) and L (for left-hand side of the scale), respectively. The modified grid form of the REP test used in this study appears as Appendix I.

Factor Analysis of the REP Test. Levy & Dugan (1965) designed a method for factor analyzing Kelly's (1955) version of the Role Construct Repertory (REP) Test in order to derive a measure of cognitive complexity and assess the patterns underlying persons' cognitive structures. This measure involves a two-stage process. During the initial phase, relevant personal constructs were elicited from each rater; the second phase involved a rating task based upon these constructs. The number of emergent factors extracted from the correlations among constructs was indicative of the raters' levels of cognitive complexity.

Each subject was presented with a 15 x 15 grid, the columns of which corresponded to role titles (e.g., father, ex-friend) representative of significant persons in the subjects' social environments. After filling in the names of those persons who corresponded to each title (no name could be used more than once), the task of construct identification began. Each row of the grid had three role-titles prespecified by circles. Constructs were generated by asking subjects to identify a characteristic that was shared by two of those persons in the triad, and yet differentiated them from the third person in that particular combination. The characteristic shared by the two individuals was used as the rating construct for that row, and was listed along with its opposite as the row heading. This process was continued until each of the fifteen rows was labeled with constructs.

These fifteen constructs, their opposites, and names of role-persons were transferred to another 15 x 15 grid. Using a 5-point (1-5) rating scale, subjects rated each of the fifteen role-persons on all of the generated constructs. In order to minimize halo errors, subjects were instructed to rate all persons on one particular construct before proceeding to the ratings on the next construct. A score at the low end of the scale (1) indicated that the construct was highly descriptive of the individual being rated; a score at the high end of the scale (5) indicated that the constructual opposite was descriptive of the ratee.

Intercorrelations among the construct ratings were factor analyzed separately for each subject, and the number of significant factors that emerged from these analyses indicated each subject's level of cognitive complexity. A fuller description of the factor-analytic scoring procedure is contained in the "Results" section. A copy of the grids used for this procedure appears as Appendix II.

Sorting Task. A method for measuring cognitive complexity that is not dependent on rating behavior was developed by Scott (1962, 1963), and is based on Goldstein and Scheerer's (1941) Object Sorting Task. This procedure involves two stages: first, the designation of significant persons in the subjects' social environments; and second, the generation of a set of relevant personal constructs.

Each subject was asked to list the names of 15 persons "who have had a significant impact on your life. These persons may be fictitious, living, or dead; they may be known to you personally or someone you 'know about.' The major criterion for choosing these persons is that they have made an important contribution to your life, as you perceive it."

Subjects were then instructed to "sort those individuals into groups based on some shared characteristic of the individuals. Persons not listed in a particular group are presumed to be persons for whom the characteristic simply does not apply, or persons who have characteristics in opposition to those listed." Groupings based upon physical characteristics (e.g., hair color, eye color, height) were to be avoided, thus ensuring that psychologically meaningful personal constructs (rather than superficial externally defined constructs) were used. Sortings were continued until the subject had exhausted the domain of constructs relevant to that group of individuals, or until a total of twenty groups had been formed.

Each subject's patterns of groupings were analyzed according to the $\underline{H}$ statistic taken from information theory: $\underline{H} = \Sigma p_{(i)}/(\log_2 1/p_{(i)})$, where $p_{(i)}$ is the proportion of individuals falling into the ith group (Attneave, 1959; Scott, 1962). This measure represents "the dispersion of persons over the set of constructs yielded by the subjects' personal construct system" (Scott, 1962, p. 408).

Rating Scale Formats

Although the rank order according to complexity of the rating scale formats used here has not been established empirically, it was hypothesized that the BARS and MSS would prove to be relatively more complex, more cognitively demanding, than the GRS or AS. This hypothesis is based on the requirement of both the BARS and MSS that raters compare scaled behavioral anchors to those behaviors of a particular ratee that are actually observed. Because the BARS requires numerical ratings on a 7-point scale, and the MSS requires ordinal ratings along a 3-point scale, the former was assumed to be the most complex scale format.

The GRS and AS were hypothesized to be less complex than the BARS and MSS, yet they were also tentatively rank ordered in terms of complexity. The GRS was presumed to be somewhat more demanding than the AS since it requires greater within-dimension discrimination in the form of a 7-point numerical scale. The AS simply requires a three-point ordinal judgement (Above Average, Average, Below Average). All four of these rating scales tapped the same nine performance dimensions, thus avoiding possible confounds resulting from unequal numbers of dimensions on the scales.

Behaviorally Anchored Rating Scale. The BARS used in this study was developed for a previous research project involving teacher evaluations (Kirkeide, Note 3). This scale was developed by undergraduate students using the method suggested by Smith and Kendall (1963). The procedure involves: (1) designation of performance dimensions and their definitions, for all important aspects of the activity to be appraised; (2) generation of behavioral statements that represent various levels of performance (superior, average, poor) for each of the dimensions; (3) retranslation of these behavioral statements by an independent group of raters in order to

determine the degree of agreement for performance dimension-behavioral statement relationships; (4) scaling of the statements by another group of raters to assess agreement on the performance levels they represent; (5) elimination of those statements that do not satisfy set criteria for agreement on the retranslation and scaling phases; and, (6) construction and testing of the scales.

Each dimension was rated separately on a 7-point linear scale, and raters were encouraged to use the behavioral statements anchoring each scale-dimension to guide their ratings. This particular BARS contained nine dimensions that corresponded to activities deemed important for success as a college-level instructor. A copy of the BARS used in this study appears as Appendix III.

Mixed Standard Rating Scale. This rating scale format, introduced by Blanz and Ghisselli (1972), was purported by them to be at least as psychometrically sound as other scales in common use. In order to develop a version of this scale, Kirkeide (Note 3) followed the initial steps in the development of the BARS outlined above. After obtaining appropriate behavioral anchors for each level of performance on the scale dimensions, three statements were chosen to represent each dimension. One of these statements reflected a superior level of performance, one moderate or average performance, and one reflected poor or inferior performance. The statements from all nine dimensions were presented, in random order, and the raters were asked to indicate if the ratee was better than, accurately described by, or worse than each behavioral statement.

Numerical ratings for each of the nine dimensions were determined by the patterns of responses given on the three associated behavioral statements.

Patterns were translated into a 7-point scale using the revised scoring procedure suggested by Saal (1979). A copy of the MSS used in this study and the key for scoring responses appear as Appendix IV.

Graphic Rating Scale. The GRS used in this study (developed by Kirkeide, Note 3) was similar in form to the BARS. The performance dimensions and their definitions are the same as those appearing on the BARS; the major difference is the absence of the behavioral anchors. This format simply presents a 7-point linear scale associated with each dimension, the low end (1) being defined as "exceptionally poor" performance and the high end (7) defined as "exceptionally good" performance. The raters were instructed to place a single mark anywhere along the line, at the point which they felt was indicative of the ratee's performance on that particular dimension. A copy of the GRS used in this study appears as Appendix V.

Alternate Rating Scale. The AS used in this study was modeled after the scale described by Schneier (1977), and was based upon the same nine performance dimensions and definitions provided in the previously described scale formats. The scale was similar to the GRS in that it contained no behavioral references associated with the dimensions. This scale, however, used a 3-point ordinal scale as the basis for ratings. Raters were instructed to place an "X" next to the term which best described the ratee's performance on each dimension: Above Average, Average, or Below Average. A copy of the AS used in this study appears as Appendix VI.

Confidence Ratings

In addition to the measures of performance that were obtained using the above rating scale formats, confidence judgements were obtained from

each rater on all rating scales. Using a 5-point Likert scale, raters were asked to indicate how certain they were that the ratings they had provided on the previous set of dimensions were an accurate reflection of the ratee's actual levels of performance. Confidence judgements were made separately for each ratee and with respect to the total set of performance ratings on each scale format (rather than separately for each dimension). Confidence scales appear in Appendices III, IV, V, and VI.

Subjects and Procedure

Raters in this study were 96 undergraduate students enrolled in General Psychology during the spring semester, 1979, who were fulfilling experimental participation requirements. Each rater was initially given the modified grid form of the REP test as a measure of cognitive complexity, followed by the task of rating three of their instructors using the 3-point alternate rating scale. The three instructors to be rated were chosen so as to be representative of a wide range of teaching performance levels; they included (1) the rater's General Psychology instructor (the same instructor for all ratees); (2) the rater's "best" instructor for that semester; and (3) the rater's "worst" instructor for that semester. If their General Psychology instructor was deemed the best or worst instructor for that semester, raters were instructed to choose the "next best" or "next worst" instructor for rating.

Following a one-week interval, participants performed the sorting task measure of cognitive complexity, and rated the same three instructors using the graphic rating scale and the mixed standard rating scale.

Another week-long interval preceded the final rating session, during which participants completed the original REP test measure of cognitive complexity and used the behaviorally anchored rating scales to rate their three instructors.

Experimental sessions were held during the thirteenth through sixteenth weeks of classes (in an 18-week semester) in order to allow raters adequate time to observe instructors' teaching styles and behavior patterns upon which ratings were based. Sessions were also arranged so as to separate, as much as possible, those rating scales and complexity measures that appear to be highly similar. For example, the modified grid form of the REP test was administered first, while the original REP test was given during the last session in order to minimize "carry-over" effects between the instruments. In addition, participants were encouraged to use each rating scale format as it was designed to be used (e.g., attending to the behavioral anchors), and not to merely attempt to duplicate the numerical ratings given at previous sessions. In this way it was expected that the differences between scale formats could be maximized.

RESULTS

In order to compare ratings obtained on the four scale formats used in this study, ratings obtained on 7-point scales (GRS, MSS, and BARS) were linearly transformed to 3-point scales. Because ratings from the AS could only be given on the basis of three points (Above Average = 3, Average = 2, Below Average = 1), these transformations allowed for direct comparisons among the ratings from different scales without omitting any of the scale values. A 3-point to 7-point transformation would necessarily preclude the use of the values 2, 3, 5 and 6 and, therefore, would underutilize the full range of the rating scale. Transformations took the following form: 1 = 1.00, 1.5 = 1.16, 2 = 1.33, 2.5 = 1.50, 3 = 1.67, 3.5 = 1.83, 4 = 2.00, 4.5 = 2.16, 5 = 2.33, 5.5 = 2.50, 6 = 2.67, 6.5 = 2.83, and 7 = 3.00.

Correlations among the mean ratings for each rater on each of the four rating scale formats appear in Table 1. These ratings are all significantly correlated, with the lowest correlation ($r(95) = .355$) being between the behaviorally anchored rating scale and the alternate rating scale. Although all of the rating scales are significantly correlated with one another, there are some differences in the ratings obtained using different rating scales; estimates of shared variance between the scales (r^2) range from 13% to 35%. These correlations lend support to the contention that raters were able to use the rating scales uniquely, possibly in the way they were designed to be used, and did not merely duplicate their ratings from one scale to the next.

Table 1

Correlations Among Rating Scale Means

	AS	MSS	GRS	BARS
AS	----	.595*	.465*	.355*
MSS		----	.557*	.570*
GRS			----	.561*
BARS				----

* $p < .001$

Cognitive Complexity and Rating Behavior

Results of the analyses of the psychometric properties of ratings obtained in this study are presented separately for each of the three complexity measures. Following a general discussion of the analysis of the measuring instrument and a description of the distribution of complexity scores that resulted from it, the psychometric properties of the ratings will be compared for high- and low-complexity raters as a function of rating scale format. Complexity scores were split at the median to form the two groups of raters.

Complexity x Ratee x Scale Format analyses of variance, with repeated measures on ratees and scale format, were used to compare rater groups with respect to leniency and halo. Complexity x Scale Format analyses of variance were used to compare range restriction for the two groups of raters. Of primary interest in these analyses were the main effects of Complexity (the degree to which complex and simple rater groups demonstrated different tendencies for each index of rating quality) and Complexity x

Scale Format interactions (the degree to which raters' cognitive complexity interacts with scale format to differentially moderate the quality of ratings for the two rater groups). Other effects not directly related to the cognitive complexity-rating scale relationship will be discussed separately, since they do not appear to affect the interpretation of the major hypotheses.

Modified Grid Form of the REP Test

The modified grid form of the REP test was analyzed according to the procedure outlined above. For every rater, ratings on a construct given to each role-person were compared to that role-person's ratings on all other constructs. A score of one (1) was given for each pair of identical ratings, and a total score, summed across the ten role-persons, determined the raters' level of cognitive complexity. Complexity scores for the 96 raters ranged from 67 (most complex) to 258 (least complex). The mean complexity score was 96.80 with a standard deviation of 23.31. Complexity scores were split at the median, 92.5, yielding two groups of 48 raters each.

Leniency. Leniency is defined as the displacement of mean ratings from the mid-point of the rating scale (Saal, Downey, and Lahey, in press). Mean ratings on each of the nine performance dimensions as a function of scale format and cognitive complexity appear in Table 2. These ratings appear to be characterized by a low to moderate degree of leniency, since all means exceed 2.0, the scale mid-point. In order to test for relative amounts of leniency, ratings were used as data points in Complexity x Ratee x Scale Format (2x3x4) analyses of variance with repeated measures on ratees and scale formats. Analyses were done separately for each of the nine performance dimensions. No significant differences (Complexity main

Table 2

Mean Ratings (Leniency) on Each Dimension as a Function of Scale Format and Complexity: Modified Grid REP Test

	Rating Scale									
	AS		GRS		MSS		BARS		Mean Ratings over all scales	
Complexity	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple
Dimension										
1	2.17	2.27	2.42	2.35	2.30	2.29	2.41	2.42	2.32	2.33
2	2.55	2.52	2.56	2.52	2.46	2.46	2.43	2.47	2.50	2.49
3	2.19	2.20	2.39	2.34	2.26	2.30	2.30	2.36	2.29	2.30
4	2.40	2.48	2.50	2.50	2.36	2.41	2.46	2.53	2.43	2.48
5	2.19	2.21	2.33	2.32	2.19	2.25	2.12	2.22	2.21	2.25
6	2.26	2.38	2.38	2.40	2.30	2.34	2.31	2.41	2.31	2.38
7	2.23	2.23	2.36	2.36	2.22	2.26	2.43	2.49	2.31	2.33
8	2.19	2.24	2.33	2.36	2.30	2.31	2.28	2.30	2.28	2.30
9	2.05	2.03	2.20	2.16	2.20	2.20	2.25	2.21	2.17	2.15
Mean	2.25	2.28	2.39	2.37	2.29	2.31	2.33	2.38	2.31	2.34

effects) were found between complex and simple rater groups on any of the nine dimensions. The *maximum* absolute difference between group means only reached .10. Comparison of mean ratings for complex and simple rater groups, collapsed across all performance dimensions, also failed to show a main effect of Complexity, $F(1,94) < 1$, n.s. ANOVA source tables appear in Appendix VII.

The interaction of cognitive complexity and rating scale format was examined to see if compatibility between raters' cognitive complexity and scale format complexity affected the relative degrees of leniency in the ratings produced by the two rater groups. No significant interactions were found on any of the nine dimensions. Collapsing these ratings across performance dimensions also failed to show a significant Complexity x Scale Format interaction, $F(1,282) = 1.30$, n.s. ANOVA source tables appear in Appendix VII.

Replication of Schneier's (1977) leniency analyses, comparison of mean ratings for complex and simple raters collapsed across performance dimensions, via *t*-tests on the AS and BARS, did not show any differences between rater groups with respect to mean ratings (Table 3). Cognitive complexity, as measured by the modified REP test, was not associated with the leniency in these ratings.

Table 3

Mean Ratings of Complex and Simple Raters on BARS and Alternate Scale

Format	Raters: Complex	Raters: Simple	df	*t*	*p*
BARS	2.33	2.38	94	-1.44	.16
AS	2.25	2.28	94	- .95	.35

Halo. Halo, the degree to which a rater is able to discriminate among different dimensions of a given ratee's performance, was assessed using the following procedure. The standard deviation of the ratings across the nine performance dimensions was calculated for each rater-ratee combination. These standard deviations were then used as data points in a Complexity x Ratee x Scale Format ANOVA with repeated measures on Ratees and Scale Format. Since the standard deviations were computed over the nine dimensions, analyses could not be performed separately for each dimension. The main effect of Complexity from this analysis was not significant, $F(1,94) = 1.02$, n.s., although there was a significant main effect of Scale Format, $F(3,282) = 25.00$, $p < .01$. The Complexity x Scale Format interaction was not significant, $F(3,282) < 1$, n.s. Mean standard deviations across dimensions appear in Table 4, and a partial source table from the ANOVA appears in Appendix VII.

Table 4

Mean Standard Deviations Across Dimensions (Halo) as a Function of Rating Scale Format and Complexity: Modified Grid REP Test

	Scale Format				
Complexity	AS	GRS	MSS	BARS	Mean
Complex	.496	.286	.314	.335	.358
Simple	.486	.267	.302	.321	.344
Mean	.491	.276	.308	.328	

Replication of Schneier's (1977) halo analyses were performed using each rater's mean standard deviations as data points in t-tests performed separately on the AS and BARS scales. Again, this analysis failed to replicate his findings. There were no significant differences in halo for

the complex and simple rater groups on either of the scale formats. The means for this analysis appear in Table 5.

Table 5

Halo of Complex and Simple Raters on BARS and Alternate Scale

Format	Raters: Complex	Raters: Simple	df	t	p
BARS	.335	.321	94	.73	.47
AS	.496	.486	94	.44	.66

Restriction of Range. Restriction of range, the degree to which raters discriminate among ratees, was investigated using each of the rater's standard deviations across ratees as data points in nine (separate analyses for each dimension) Complexity x Scale Format (2x4) ANOVAs with repeated measures on scale format. Because the standard deviations were computed over ratees, they (ratees) could not be analyzed as a separate effect. Mean standard deviations as a function of Complexity and Scale Format appear in Table 6. Although only one of the main effects of Complexity proved to be significant (Dimension 4, Instructor Knowledge, $F(1,94) = 5.07$, $p < .05$) with high complexity raters showing less range restriction (larger standard deviations), this trend can also be seen for a number of other dimensions. Repeating this analysis with the data collapsed across the nine dimensions showed a significant Complexity main effect, $F(1,94) = 4.08$, $p < .05$. This indicates that over all the rating dimensions and over all the scale formats, complex raters discriminated more among the ratees than did the simple raters.

Table 6

Mean Standard Deviations Across Ratees (Range Restriction) as a Function of Scale Format and Cognitive Complexity: Modified Grid REP Test

	Rating Scale									
	AS		GRS		MSS		BARS		Mean Ratings over all scales	
Complexity	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple
Dimension										
1	.802	.713	.441	.406	.473	.378	.446	.401	.541	.475
2	.676	.639	.447	.365	.526	.433	.485	.383	.534	.455
3	.641	.669	.414	.427	.388	.384	.469	.474	.478	.489
4	.747	.627	.422	.339	.473	.372	.431	.368	.519	.427[a]
5	.798	.744	.588	.478	.473	.499	.522	.544	.595	.566
6	.684	.603	.507	.379	.403	.409	.443	.388	.509	.445
7	.722	.632	.507	.413	.396	.373	.470	.388	.524	.451
8	.743	.663	.534	.449	.444	.394	.450	.372	.543	.469
9	.653	.629	.556	.485	.361	.334	.501	.501	.518	.488
Mean	.718	.658	.491	.416	.438	.397	.469	.424	.529	.474[a]

[a] Main effect for Complexity $p < .05$.

Complexity x Scale Format interactions were not significant for any of the nine dimensions individually, and no significant interaction emerged when ratings were collapsed across the dimensions, although a significant main effect of Rating Scale Format was found in all analyses. ANOVA source tables for these range restriction analyses appear in Appendix VII.

Separate analyses, using t-tests between rater groups, were performed for the BARS and AS in order to replicate Schneier's (1977) analyses. These analyses were done on the standard deviations collapsed over dimensions. Results appear in Table 7. No significant differences among the two groups of raters was found for the BARS; however, a "marginal" difference between the groups emerged on the AS. In general, therefore, the complex and simple rater groups did not differ in restriction of range.

Table 7

Restriction of Range of Complex and Simple Raters on BARS and Alternate Scale

Format	Raters		df	t	p
	Complex	Simple			
BARS	.469	.424	94	1.41	.16
AS	.718	.658	94	1.83	.07

Overall, the results presented here do not replicate those of Schneier (1977) and do not support his compatibility theory of rating behavior. As measured by the modified grid form of the REP test, cognitive complexity was not shown to be a moderator of rating quality.

Factor Analysis of the REP Test

Each subject's REP test ratings were factor analyzed according to the procedure suggested by Levy and Dugan (1956) to yield a measure of cognitive complexity. The factor analytic procedure they suggested was Thurstone's multiple group method with orthogonal rotation of the extracted factors. Another factor analysis procedure, the principal-axes factor analysis, was used in this study, however, so that the analyses could be handled by the computer. Thurstone's multiple group method is a short-cut procedure which reduced the tedious arithmetic associated with factor analysis prior to the advent of the electronic computer. It has been shown (Overall & Klett, 1972, p. 152) that these two procedures yield highly similar factor solutions, so no significant deviations from Levy and Dugan's original intent were anticipated.

Following completion of the REP test outlined above, each subject's pattern of personal construct correlations was factor analyzed. The number of significant emergent factors for each subject was taken as their measure of cognitive complexity. Significance of the factors was determined by counting only those factors that had eigenvalues in excess of the average initial communality estimate for all factors. Because the majority of the correlation matrices for the 96 raters were singular (preventint the calculation of $\underline{R}^2$), initial estimates of communality were the maximum off-diagonal elements of the correlation matrix for each construct. Average correlations (communalities) were calculated using Fisher's $\underline{r}$ to $\underline{z}$ transformation. Factors whose eigenvalues were less than this average communality estimate were not included.

Complexity scores (number of factors) for the 96 raters ranged from 2 (least complex) to 8 (most complex), with a mean of 4.47 and a standard

deviation of 1.15. Complexity scores were split at the median (median = 4.5) to yield two groups of raters. The simple (low-complexity) rater group included 49 raters (in order to take into account all raters with 2, 3, or 4 emergent factors) and the complex rater group consisted of 47 raters (all raters with 5, 6, 7, or 8 emergent factors). All analyses of variance used to compare ratings for these groups were unequal n ANOVAs.

Leniency. Ratings on each of the nine performance dimensions were the data points in a set of Complexity x Ratee x Scale Format ANOVAs with repeated measures on ratees and scale formats. Examinations of the main effects of Complexity (collapsed across scale formats) on all dimensions showed no significant effects. (Source tables for this ANOVA appear in Appendix VIII.) Mean ratings as a function of scale format and the factor-analytic measure of cognitive complexity appear in Table 8. Again, all means are in excess of the scale mid-point, indicating an overall tendency on the part of the raters to give lenient ratings, with no differential tendencies for the complex and simple rater groups. In addition, Complexity and Scale Format did not interact on any of the dimensions to affect the mean rating levels.

When ratings were combined over all dimensions, leniency analyses were unchanged. A Complexity x Ratee x Scale Format ANOVA using scale means as data points failed to demonstrate a significant Complexity main effect, $F(1,94) < 1$, n.s., and showed no significant Complexity x Scale Format interaction, $F(3,282) < 1$, n.s.

Halo. Standard deviations of each rater's ratings across the nine performance dimensions were the data points in a Complexity x Ratee x Scale Format ANOVA. No significant main effect of Complexity was found, $F(1,94)$

Table 8

Mean Ratings (Leniency) on Each Dimension as a Function of Scale Format and Cognitive Complexity: Factor Analysis

	Rating Scale									
	AS		GRS		MSS		BARS		Mean Ratings over all scales	
Complexity	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple
Dimension										
1	2.24	2.19	2.39	2.38	2.30	2.29	2.37	2.44	2.32	2.33
2	2.51	2.56	2.52	2.56	2.46	2.47	2.41	2.47	2.48	2.51
3	2.22	2.17	2.37	2.36	2.30	2.27	2.34	2.30	2.31	2.28
4	2.41	2.47	2.52	2.48	2.37	2.41	2.47	2.50	2.44	2.46
5	2.22	2.17	2.34	2.31	2.25	2.19	2.19	2.14	2.25	2.20
6	2.33	2.31	2.39	2.39	2.30	2.34	2.32	2.37	2.33	2.35
7	2.22	2.24	2.34	2.38	2.26	2.22	2.44	2.46	2.31	2.33
8	2.19	2.23	2.34	2.35	2.30	2.31	2.32	2.24	2.29	2.28
9	2.05	2.03	2.17	2.19	2.22	2.18	2.22	2.22	2.16	2.16
Mean	2.27	2.27	2.38	2.38	2.30	2.30	2.36	2.35	2.33	2.32

< 1, n.s., indicating similar standard deviations across dimensions for both rater groups. No significant Complexity x Scale Format interaction emerged, $F(3,282) = 1.05$, n.s., showing no differential effects of cognitive complexity as a function of rating scale format. A significant main effect of Rating Scale Format did emerge, however, $F(3,282) = 17.91$, $p < .01$. Mean standard deviations appear in Table 9. Source Tables appear in Appendix VIII.

Table 9

Mean Standard Deviations Across Dimensions (Halo) as a Function of Rating Scale Format and Complexity: Factor Analysis

	Scale Format				
Complexity	AS	GRS	MSS	BARS	Mean
Complex	.493	.284	.310	.318	.351
Simple	.488	.269	.306	.338	.350
Mean	.490	.276	.308	.328	

Restriction of Range. Standard deviations across ratees for each rater were used as data points in Complexity x Scale Format ANOVAs on each dimension, with repeated measures on scale formats. Main effects of Complexity reached significance for only one performance dimension (Dimension 8, Student-Teacher Relations, $F(1,94) = 4.53$, $p < .05$) with complex raters showing less range restriction (larger standard deviations) than simple raters. This trend is apparent, though not significant, for a number of other dimensions. Combining ratings over all dimensions, however, failed to produce a significant main effect of Complexity ($F(1,94) = 2.16$, n.s.).

Table 10

Mean Standard Deviations Across Ratees (Range Restriction) as a Function of Scale Format and Cognitive Complexity: Factor Analysis

	Rating Scale									
	AS		GRS		MSS		BARS		Mean Ratings over all scales	
Complexity	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple
Dimension										
1	.746	.769	.436	.410	.473	.378	.467	.380	.531	.485
2	.681	.633	.419	.393	.481	.479	.448	.420	.507	.481
3	.684	.627	.450	.391	.398	.374	.482	.460	.503	.463
4	.717	.657	.373	.389	.447	.399	.409	.390	.486	.459
5	.759	.782	.518	.548	.530	.443	.501	.564	.577	.584
6	.699	.588	.458	.428	.456	.356	.447	.384	.515	.439
7	.695	.659	.496	.424	.367	.401	.427	.432	.496	.479
8	.758	.648	.539	.445	.468	.370	.437	.385	.550	.462[a]
9	.648	.635	.530	.512	.398	.297	.534	.467	.528	.478
Mean	.710	.666	.469	.438	.446	.389	.461	.431	.522	.481

[a] Complexity main effect $p < .05$.

No significant Complexity x Scale Format interaction effects were found for any of the nine performance dimensions individually, nor for the analysis combined across dimensions (source tables appear in Appendix VIII). Mean standard deviations as a function of Scale Format and Complexity appear in Table 10. A significant main effect of Scale Format was found for each of these analyses with the AS yielding larger standard deviations than any of the other scales (which do not differ from one another).

Sorting Task

Following Scott's (1962, 1963) procedure, groupings of individuals obtained from the sorting task measure of cognitive complexity described above were analyzed using the H statistic. For each rater, a complexity score was derived using the formula: $H = \Sigma p_{(i)}/(1/\log p_{(i)})$, where $p_{(i)}$ is equal to the proportion of individuals falling into the ith group. Complexity scores for the 96 raters ranged from 0 (least complex) to 3.373 (most complex), with a mean of 2.468 and a standard deviation of .529. Splitting scores at the median, 2.560, yielded two groups of 48 raters.

Leniency. Results of the nine Complexity x Ratee x Scale Format ANOVAs (repeated over ratees and scale formats) again failed to show any significant differences between complex and simple rater groups in absolute levels of leniency collapsed over scale formats (Complexity main effects), or in tendencies toward lenient ratings as a function of scale format (Complexity x Scale Format interactions). Mean ratings appear in Table 11. Combining ratings over all dimensions did not alter this result. Cognitive complexity as measured by the sorting task was not related to leniency in the ratings. Source tables for these analyses appear in Appendix IX.

Table 11

Mean Ratings (Leniency) on Each Dimension as a Function of Scale Format and Cognitive Complexity: Sorting Task

	Rating Scale									
	AS		GRS		MSS		BARS		Mean Ratings over all scales	
Complexity	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple
Dimension										
1	2.23	2.21	2.43	2.33	2.31	2.27	2.44	2.39	2.35	2.30
2	2.57	2.50	2.56	2.52	2.49	2.44	2.46	2.44	2.52	2.47
3	2.23	2.17	2.39	2.35	2.27	2.30	2.37	2.30	2.31	2.28
4	2.41	2.47	2.51	2.50	2.40	2.38	2.49	2.50	2.45	2.46
5	2.22	2.17	2.33	2.33	2.19	2.25	2.17	2.17	2.23	2.30
6	2.38	2.26	2.40	2.38	2.32	2.32	2.34	2.37	2.36	2.33
7	2.24	2.22	2.36	2.36	2.21	2.27	2.47	2.45	2.32	2.32
8	2.19	2.23	2.33	2.37	2.29	2.32	2.36	2.22	2.29	2.28
9	2.01	2.08	2.18	2.18	2.22	2.18	2.24	2.22	2.16	2.16
Mean	2.28	2.26	2.39	2.37	2.30	2.30	2.37	2.34	2.33	2.32

Halo. A Complexity x Ratee x Scale Format ANOVA, using standard deviations across dimensions as data points, was again used to supply halo information. No significant Complexity main effect ($F(1,94) = 1.07$, n.s.) or Complexity x Scale Format interaction ($F(3,282) = 1.11$, n.s.) emerged. A significant Scale Format main effect was found, with the AS yielding the largest standard deviation. Mean standard deviations as a function of Complexity and Scale Format appear in Table 12.

Table 12

Mean Standard Deviations Across Dimensions (Halo) as a Function of Rating Scale Format and Complexity: Sorting Task

	Scale Format				
Complexity	AS	GRS	MSS	BARS	Mean
Complex	.503	.284	.320	.323	.358
Simple	.478	.268	.296	.332	.344
Mean	.490	.276	.308	.328	

Restriction of Range. Complexity x Scale Format ANOVAs using standard deviations across ratees for each rater as data points were done separately for each of the nine dimensions. Although none of these analyses proved to be statistically significant (source tables appear in Appendix IX), a trend similar to that found with the other complexity measuring techniques can be seen in the means presented in Table 13. Cognitively complex raters tended to provide ratings with greater dispersion among the ratees; however, this effect is less than overpowering. When ratings were combined over dimensions, no significant main effect for Complexity was found. A significant Scale Format effect did emerge in all analyses.

Table 13

Mean Standard Deviations Across Ratees (Range Restriction) as a Function of Scale Format and Cognitive Complexity: Sorting Task

	Rating Scale									
	AS		GRS		MSS		BARS		Mean Ratings over all scales	
Complexity	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Simple
Dimension										
1	.776	.740	.417	.430	.456	.395	.433	.415	.520	.495
2	.658	.656	.393	.419	.475	.484	.444	.424	.492	.496
3	.693	.618	.427	.414	.385	.388	.474	.468	.495	.472
4	.708	.666	.404	.358	.419	.427	.410	.389	.485	.460
5	.800	.741	.556	.510	.453	.520	.497	.569	.576	.585
6	.702	.584	.467	.419	.412	.400	.422	.409	.501	.453
7	.701	.653	.501	.419	.382	.386	.427	.431	.503	.472
8	.762	.644	.523	.461	.456	.382	.430	.392	.543	.470
9	.677	.605	.567	.475	.333	.362	.517	.485	.524	.482
Mean	.720	.657	.473	.434	.419	.416	.450	.443	.515	.487

In addition, there were no significant Complexity x Scale Format interactions for any of the individual dimensions analyses, nor for the combined analysis (see Appendix IX for source tables).

Other ANOVA Effects

The ANOVA effects that are not directly related to the cognitive complexity-rating scale format relationship were investigated in order to determine if subjects were providing ratings in accordance with the instructions administered at each rating session. It has already been mentioned that, based on the correlations among the rating scales, subjects seemed able to use the rating scales in different ways (and perhaps even in the way they were designed to be used), and were not merely duplicating their ratings from one session to the next. Differences between ratings obtained by the different scales were examined by looking at the main effects for Scale Format in each analysis. The Ratee main effects were also of interest since raters were specifically asked to choose, as ratees, instructors whose levels of performance differed widely (i.e., their "best" and "worst" instructors).

Analyses for mean rating levels (leniency) tended to show significant main effects for rating scale format on individual dimensions and when ratings were combined over dimensions. Post-hoc comparisons of the means (Newman-Keuls Tests) showed that the mean ratings provided on the AS were significantly smaller (less lenient) than the means for the other three scale formats, and that the mean ratings for the MSS tended to be smaller than the mean ratings provided by the BARS and GRS. BARS and GRS ratings did not differ significantly from one another. This effect may be explained, in part, as a function of the scaling procedures. Ratings on the AS could

be given at only one of three points (1, 2 or 3) and no intermediate ratings could be given; whereas, up to five ratings between these whole numbers were possible on the other three scale formats. Raters were instructed to rate persons whose levels of performance varied widely (their "best" and "worst" instructors), which would tend to spread the ratings over the entire range of the scale. In order to spread the ratings out, however, the mean for the AS would have to reflect a lower overall level than would be the case for the GRS, MSS, or BARS where intermediate ratings were possible.

As expected, a significant Ratee main effect was found on all dimensions and analyses combined over dimensions. Because raters were asked to rate their "best" and "worst" instructors, it was anticipated that ratings would differ for different ratees. Post-hoc comparisons showed that the ratings for the extreme instructors did, indeed, differ in the expected direction, and that ratings for the General Psychology instructor fell somewhere between these. Ratings for the General Psychology instructor did not always differ significantly from the ratings for the "best" instructor.

Halo and range restriction analyses, which were based on standard deviations of the ratings, consistently showed a main effect of Scale Format. Post-hoc comparisons (Newman-Keuls Tests) showed that, in all cases, the largest standard deviations were associated with the alternate scale. These effects may also be explained as a result of the scaling procedures. Since no ratings between 1, 2 or 3 could be given (and up to 5 ratings between these values were possible for the other three scales), the standard deviations for the AS are somewhat inflated. This explanation is, admittedly, post-hoc and other reasonable hypotheses cannot be totally ruled out on the basis of these data.

Ratee main effects were examined with respect to the halo analyses. Since standard deviations across Ratees was the measure of range restriction, they could not be analyzed as separate effects. For all three complexity measures, halo analyses produced a significant Ratee main effect (see Appendix X). Post-hoc comparisons of these effects showed that, in all cases, the largest standard deviations were associated with the "worst" instructor and the smallest standard deviations were associated with the "best" instructor. The ratings for the intermediate instructor were characterized by an intermediate level of dispersion.

These results may be explained by the "differential accuracy phenomenon" (Gordon, 1970, 1972; Landy & Farr, 1980). It has been reported that ratings are often "more accurate" when the behavior in question is favorable rather than unfavorable. In addition, Baker & Shuck (1975) found that this phenomenon appears to be limited to only certain rating dimensions and not to others. Although the exact reasons for this phenomenon are unclear, the results of this study seem to corroborate earlier findings.

Confidence in Ratings

Complexity x Ratee x Scale Format ANOVAs were used to analyze confidence ratings. Mean confidence judgements as a function of Complexity and Rating Scale Format appear in Table 14. Because there were no differences in these results as a function of the measuring technique for cognitive complexity, the results are presented for one measure only, the modified grid REP test.

Table 14

Mean Confidence Ratings as a Function of Rating Scale Format and Cognitive Complexity: Modified Grid REP Test

Complexity	Rating Scale AS	GRS	MSS	BARS	Mean
Complex	3.99	4.05	4.13	4.19	4.09
Simple	4.07	4.13	4.13	4.17	4.13

Responses to this measure were highly inflated with an overall mean of 4.11 and a standard deviation of .586. No significant differences were found for Complexity (complexity main effect $F(1,282) < 1$, n.s.) or for Scale Format, $F(3,282) = 1.84$, n.s. In addition, no Complexity x Scale Format interactions were found, $F(3,282) = 1.33$, n.s.

Significant Ratee main effects (for all three measures of complexity) were found, and this is also consistent with the differential accuracy phenomenon hypothesis. Raters were most confident in their ratings of their "best" instructor and were least confident in their ratings of their "worst" instructor.

Relationships Among Measures of Cognitive Complexity

In order to measure the relationships among the three measures of cognitive complexity used in this study, three bivariate Pearson-Product-Moment correlations were computed. These correlations appear in Table 15. Because the modified grid form of the REP test scores are inversely related to the level of cognitive complexity (lower scores indicate higher complexity levels), the negative correlations found here are indicative of congruent relationships among these measures.

Table 15

Correlations Among Measures of Cognitive Complexity

	Modified REP Test	Factor Analysis	Sorting Task
Modified REP Test	-----	-.196[a]	-.365[b]
Factor Analysis		-----	.246[c]

[a] $p = .056$ [b] $p < .01$ [c] $p < .05$

Results of these bivariate correlations indicate that the three measures of cognitive complexity used in this study at least partially measure the same construct; however, the percent of shared variance among the measures is small. Estimates of shared variance (r^2) range from 4% for the modified grid and factor-analytic measures to 13% for the modified grid REP test and the sorting task measures. This indicates that a substantial portion of the variance in these measures is unaccounted for by the variance in other measures.

To test the possibility that the sorting task and factor analysis measures account for separate portions of the variance in the modified REP test, a multiple regression coefficient was calculated using the modified REP test as the dependent variable and the other two measures as independent variables. The regression coefficient resulting from this analysis, $R = .381$, does not differ from the bivariate correlation of the modified REP test and the sorting task measure; the proportion of variance added by the factor analysis measure is not significant.

Because analyses of the ratings from this study were based on median splits with respect to the complexity measures and the absolute values of the scores were categorized into dichotomous (complex or simple) groups, contingency tables were constructed in order to assess the consistency with

which the same raters fell into the simple and complex rater groups across the three measures of cognitive complexity. Chi-square associations among the measures appear in Table 16.

Table 16

Chi-Square Values for Contingencies
Between Measures of Cognitive Complexity

	Modified REP Test	Factor Analysis	Sorting Task
Modified REP Test	-----	.167	2.667
Factor Analysis		-----	4.667*

* $p < .05$

These analyses indicate that raters who are classified as complex on one measure of cognitive complexity are not any more likely than the simple raters to be classified as complex on another measure of cognitive complexity; the possible exception to this is the relationship between the factor analysis classification and the sorting task classification. It should be noted that the chi-square value reported for these two measures only reflects the congruent classification of one more individual than is classified by the sorting task and modified REP test measures which was not statistically significant.

Taken in conjunction with the correlations previously reported, these results seem to indicate weak or insignificant relationships among the various measures of cognitive complexity. It appears, therefore, that reliance on a single index (or measuring technique) as an accurate reflection of a person's cognitive complexity may provide an incomplete, and perhaps biased, picture.

Summary

The results of this study failed to support the hypothesis that cognitive complexity is a moderator of rating behavior. With few exceptions, comparisons between cognitively-complex and cognitively-simple rater groups showed no systematic differences in their tendencies to provide ratings characterized by leniency, halo, or restriction of range. These results (or lack thereof) were consistent, regardless of the type of rating scale format used and were also consistent across three ostensibly different techniques for measuring cognitive complexity.

All of the analyses reported here were repeated using the upper, middle and lower thirds of the complexity scores to split the raters into three complexity groups--complex, moderate, and simple. Although a few of the separate dimension effects for Complexity and Complexity x Scale Format analyses reached significance with this procedure, their numbers did not appear great enough to warrant any changes in the conclusions. Using the extremes of the complexity scores did not change the overall pattern of the relationships between cognitive complexity and rating scale formats that were found by using two rater groups. In addition, analyses were repeated on the GRS, MSS and BARS untransformed ratings (based on their original 7-point scales) with no differences in results.

DISCUSSION

Research in the area of performance appraisal systems in recent years has begun to adopt an "interactionist" or process approach. Because investigations that were limited to the characteristics of rating scale formats have proven incapable of explaining the properties of ratings, attention is now being directed toward rater characteristics which may play a role in moderating rating behavior. In particular, there has been a call for the investigation of raters' cognitive processes as a way of increasing our understanding of biases in performance appraisal ratings (Landy & Farr, 1980).

In what has become a widely cited study, Schneier (1977) proposed a compatibility theory of performance appraisal. Cognitive complexity, the degree to which a person is able to discriminate among dimensions of complex stimuli, was found to interact with rating scale complexity to reduce leniency and restriction of range in ratings. In addition, cognitively complex raters were found to provide ratings characterized by lower levels of halo, regardless of the rating scale used. The present study was designed to further investigate the relationship between cognitive complexity and the use of different types of rating scale formats in order to determine the dimensions of both that may influence the rating process.

In contrast to Schneier's (1977) findings, this study failed to demonstrate any convincing relationship between cognitive complexity and rating behavior. Cognitively complex raters did not differ from cognitively simple raters when their ratings were compared with respect to leniency and halo. Although a weak trend toward differences between complex and simple raters was found with respect to restriction of range, this effect only reached significance for one of three measures of cognitive complexity used in this study.

Of greater importance, however, was the failure of any significant Complexity x Scale Format interactions to emerge. Ratings obtained when raters' cognitive complexity and the presumed complexity of the rating scale format were congruent did not differ from ratings obtained when rater complexity and scale format complexity were not compatible. This absence of significant interactions on all indices of rating quality casts doubt on the generalizability of Schneier's (1977) original compatibility hypothesis. Possible explanations for the lack of congruence between the findings of these two studies are discussed as a function of the rating scale composition, of the rater samples, and, finally, with respect to the concept of cognitive complexity itself.

Rating Scale Composition

Schneier (1977) proposed that the complexity of a rating scale is a function of the number of discriminations that must be made between dimensions of performance and the number and quality of discriminations required on each individual dimension. He argued that behaviorally based rating scales, which require raters to compare scaled behavioral anchors to actual observed behavior and that further necessitate ratings on a 7-point numerical continuum, are more complex than formats that do not require behavioral comparisons and contain "ordinal" scales with three rating options. His results with respect to this contention, however, are confounded, since the performance appraisal formats he employed varied both in the number of dimensions tapped and the type of scale used for judging each dimension.

The rating formats used in this study each tapped the same number of performance dimensions (nine), although the types of scales used to obtain judgements on these dimensions did differ substantially from one scale format

to the next. Two of the scale formats required numerical ratings on continuous scales, while the other formats used only three options, ordinally arranged, as the basis for rating. Two of the formats used behavioral anchors to define levels of performance on each dimension, and two were ambiguous in that raters were allowed to interpret the exact meaning of performance levels idiosyncratically. Because no differences were found between complex and simple rater groups as a function of these different scaling techniques, and minimal differences were found among the scales themselves, there is no reason to assume that any of the scales were more or less complex than any of the others. This does not, however, preclude the possibility that rating scale formats are differentially complex; in fact, this notion remains intuitively appealing. It can be safely said that differences in the scales that may exist did not interact with cognitive complexity to affect the ratings obtained in this study.

Another argument that addresses itself to the complexity of the rating scales and that may account for the failure of this study to replicate Schneier's (1977) work is based on differences in the BARS formats used in the two studies. Since the "simple" scales did not appear to differ substantially--Schneier's (1977) AS tapped ten dimensions using three rating choices and this study used a nine-dimension AS with the same three options--it is unlikely that these two scales were differentially complex. Differences in the BARS, on the other hand, can be noted. Schneier's 14-dimension BARS contained 183 behavioral reference statements in contrast to the nine-dimension BARS used here, which contained only 50 behavioral anchors. Given these differences in composition, it is possible that the BARS were not equally cognitively demanding.

It may be the case that the BARS used by Schneier (1977) were highly confusing (too many dimensions and too many behavioral anchors), thereby over-extending the relatively limited discriminatory abilities of the cognitively simple raters. Cognitively simple raters are able to discriminate among dimensions of stimuli, but to a lesser extent than cognitively complex raters. The BARS used by Schneier may have exceeded that capacity, producing ratings that were characterized by different levels of psychometric "errors."

Bernardin and Boetcher (Note 2) used a similar explanation to account for their inability to replicate Schneier's (1977) findings. Using only one rating scale format, a seven-dimension behaviorally based scale, they found that cognitively simple raters did not differ from complex raters with respect to leniency or halo. They argued that differences in BARS developmental procedures can often render these scales differentially confusing. Special care taken during the retranslation and scaling phases of BARS development, therefore, may minimize the cognitive processing required of the raters, and thereby minimize tendencies toward rating "errors." If the rating scales can be developed so as not to overly tax the discriminatory abilities of the raters, particularly the less complex raters, differences in rating quality as a function of interactions between cognitive complexity and rating scale complexity may also be minimized.

Differences in rating scale composition may also account for Schneier's (1977) finding that the BARS and AS had different levels of "operational utility" for complex and simple raters. He found that complex raters were more confident with, more satisfied with, and preferred the complex BARS and that the simple raters preferred, were more confident with, and were more satisfied with the simple AS. Unfortunately, it is impossible to determine whether these judgements were a function of the total rating scale

complexity (number of dimensions and scales used) or solely a function of the number of performance dimensions tapped. Raters in the present study were equally confident in their use of all four rating scale formats. Because these scales were composed of identical dimensions, it appears as if Schneier's operational utility findings were more a function of the number of dimensions (or the kinds of dimensions) than the type of scale used to gather the information on each dimension. If this is the case, increased attention may be needed during the dimension identification and definition phase of BARS development in order to accurately reflect the aspects of performance that are important to the appraisal process.

It appears, therefore, that the number of dimensions tapped by "complex" rating scales, and the procedures used in their development, may be more important indices of the complexity of the cognitive demands required of the raters than the types of scales used to record the judgements. Further investigations of this hypothesis appear warranted; however, future investigations need to pay special attention to potential order effects of rating scale administration, a variable uncontrolled in the present study.

Rater Samples

The possibility that the raters in the present study were different with regard to their levels of cognitive complexity than the raters in Schneier's (1977) study was investigated. Because Schneier used unskilled manufacturing workers as raters, and the present study (similar to Bernardin and Boetcher, Note 2) used student raters, the lack of consistency in the findings would be more comprehensible if it could be shown that levels of cognitive complexity differed across these samples. Also, convention (Johnson & Centers, 1973) dictated that complexity scores be split at the median to obtain complex and simple sub-groups, a procedure that could

result in totally different levels of complexity being classified as complex/simple depending on the distribution of scores in different samples. Comparative data were only available for one of the complexity measures used in the present study--the modified grid form of the REP test.

The range of possible scores on this form of the REP test is 30-450. The range of scores for the present sample was 67 to 258, and the range for Schneier's (1977) unskilled worker sample was reported to be 65 to 232. Medians for the two samples were 92 and 92.5, respectively. Given these results, there is no reason to assume that differential results bearing on the relationship between cognitive complexity and rating behavior are a function of differences in complexity scores. In fact, the relative ranges and medians of the two samples' scores imply little or no cognitive complexity differences between these samples.

This comparison is consistent with other comparisons of complexity scores obtained from different samples. Such characteristics as sex, level in the organizational hierarchy, occupation (student vs. non-student), year in college, and college major were not found to alter distributions of scores on the modified grid form of the REP test (Schneier, 1979). In addition, test-retest reliability coefficients for this measure across various samples consistently range from $\underline{r}$ = .60 to $\underline{r}$ = .90 (Tripodi & Bieri, 1963; Schneier, 1979). Thus, the procedure of splitting complexity scores at the median seems to result in complex and simple sub-groups that are comparable to sub-groups in other samples, at least with respect to this measure of complexity.

Normative and reliability data for the factor-analytic measure and the sorting task measure of cognitive complexity are not available. Studies that have used these measures rarely report the descriptive statistics that are

necessary for comparing complexity scores across samples. In order to illustrate their analyses, Levy and Dugan (1956) did report the factor structures for two subjects (admittedly not a strong basis for comparison), both of which yielded four significant emergent factors on the REP test. Although this comparison is crude at best, four factors were found to be the mode in the present study; 28 of 96 raters had four factors, indicating that this set of complexity scores may have some consistency with other samples. In addition, Schneier (1979) correlated the modified grid form of the REP test with a sorting task measure of complexity and found a correlation of r (175) = -.19, $p < .05$. This correlation is at least not inconsistent with the relationship found here (r (95) = -.37, $p < .01$). Without more convincing data, however, it is impossible to tell if cognitive complexity, as measured by the factor-analytic procedure and the sorting task, is reliable across samples.

Cognitive Complexity

In spite of the stability of the modified grid form of the REP test across samples and time, the lack of strong relationships between it and other measures of cognitive complexity raises serious questions concerning the comparability of studies using different measuring strategies. The three measures of cognitive complexity, while significantly correlated in the expected direction, did not share substantial portions of their variances. Given this state of affairs, it is not possible to determine the nature of the skills and abilities that comprise complex (as opposed to simple) raters. Also, the failure to find any relationship between cognitive complexity and rating behavior, regardless of the measuring instruments or rating scale formats used, makes it impossible to determine which aspects of cognitive complexity may have been responsible for Schneier's (1977) findings.

The three measuring techniques used to assess cognitive complexity in this study were presumed to vary along at least two dimensions. Only two of the procedures, the modified grid form of the REP test and the factor analysis of the original REP test, relied directly on rating behavior as an index of cognitive complexity. These two measures were only "marginally" correlated (r (95) = -.196, p = .06), however, indicating that rating behavior *per se* may not be a particularly salient feature of cognitive complexity. The results of this study, in contrast to Schneier's (1977) findings, suggest that this type of rating task is unrelated to rating in the context of performance appraisal. Neither of the two versions of the REP test was found to be a predictor of rating quality in this study.

Two of the measures of cognitive complexity were dependent on the subjects' use of their own personal constructs, rather than constructs provided for them. The factor analysis measure of the REP test and the sorting task measure of cognitive complexity both utilize constructs that are elicited by each subject individually, and the context for the generation of the constructs was presumed to be similar (significant persons in the subjects' social environments). Scores obtained on these two measures were significantly correlated, yet the percentage of shared variance in the measures is extremely small, r (95) = -.246, $p < .05$; $r^2 = .06$. Given the importance of personal constructs to the concept of cognitive complexity, this relationship seems unusually small, indicating that cognitive complexity, whatever it may be, is more than a reflection of the contents of individual's personal construct systems.

Past research has concluded that no differences exist in cognitive complexity scores as a function of constructs. Tripodi & Bieri (1963) found that the distributions of complexity scores when subjects used their

own personal constructs were not different from the distributions of scores when constructs were provided for them; they also found significant rank-order correlations between scores based on a grid task with provided constructs and a grid task using "own" constructs. Similar results have been reported by Metcalfe (1974) and by Adams-Webber (1970). In spite of these findings, the two REP test grid measures used in this study are only marginally related to one another. This relationship is not totally inconsistent with those reported by Metcalfe (1974) and Adams-Webber (1970), both of whom highlighted the low magnitudes of their reported correlations (r's in the vicinity of .33), and cautioned against over-reliance on the similarity of the measures. Both of these authors advocated the use of "own" constructs for deriving measures of complexity when available.

Perhaps the most surprising result to come from the examination of these three measures of cognitive complexity is the fact that the two measures which seemed, on the surface, to have the least in common--the modified grid form of the REP test and the sorting task--proved to be most strongly related (r (95) = -.365, p < .01). One of these measures relies on the use of provided constructs, whereas the other uses constructs supplied by the subjects; one measure uses a sorting task measure and the other a rating task. Their commonality results from the fact that they share the same theoretical base (construct validity); both are measures of cognitive complexity. This does not, however, help to explain the specific skills or abilities that are related to or comprise cognitive complexity, and their lack of relationship to rating behavior, in the context of the present study, fails to provide any insight into the nature of the cognitive complexity-rating behavior relationship. The specific nature of this relationship, if it does exist, simply cannot be determined from this study.

Conclusions

While the definition of cognitive complexity bears some conceptual similarity to the cognitive processes involved in performance appraisal (ratings), problems result from our apparent inability to clearly measure a person's capacity for perceiving behavior in a multi-dimensional way. Further problems emerge when we attempt to identify the characteristics of rating scales that seem to require different levels of this complex ability. These difficulties, coupled with (and probably contributing to) the lack of consistent findings regarding the relationship between cognitive complexity and rating scale formats, demand, at the very least, that caution be exercised before "jumping on the cognitive complexity bandwagon." Without further support for the presence of a reliable relationship between cognitive- and rating scale-complexity, in terms of both the psychometric properties of ratings and the operational utility of the rating scales, there seems to be little reason to believe that cognitive complexity and rating scale formats are related in any systematic way.

In spite of the fact that cognitive complexity was not found to be a predictor of rating behavior in the present study, the continued exploration into the cognitive processes of raters and their relationship to other aspects of the performance appraisal system remains an important area of research. Examinations of the interactive phases of the rating process, as advocated by Landy & Farr (1980), appear to be the most valuable route toward insight into the quality and the utility of performance appraisal data.

REFERENCE NOTES

1. Schneier, C. E. Personal communication, July 11, 1979.

2. Bernardin, H. J., & Boetcher, R. The effects of rater training and cognitive complexity on psychometric error in ratings. Unpublished manuscript, Old Dominion University, 1979.

3. Kirkeide, L. K. Rating scale format and the effectiveness of training raters to minimize rating errors. Unpublished manuscript, Kansas State University, 1979.

REFERENCES

Adams-Webber, J. R. Cognitive complexity and sociality. British Journal of Social and Clinical Psychology, 1969, 8, 211-216.

Adams-Webber, J. R. Elicited versus provided constructs in repertory grid technique: A review. British Journal of Medical Psychology, 1970, 43, 349-354.

Adams-Webber, J. R. The complexity of the target as a factor in interpersonal judgement. Social Behavior and Personality, 1973, 1, 35-38.

Adams-Webber, J. R. Personal construct theory: Concepts and applications. New York: John Wiley and Sons, 1979.

Attneave, F. Applications of information theory to psychology. New York: Holt-Dryden, 1959.

Bannister, D. Conceptual structure in thought disordered schizophrenics. Journal of Mental Science, 1960, 106, 1230-1249.

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

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 1976, 61, 75-79.

Bernardin, H. J., & Walter, C. S. The effects of rater training and diary keeping on psychometric errors in rating. Journal of Applied Psychology, 1977, 62, 64-69.

Bieri, J. Cognitive complexity-simplicity and predictive behavior. Journal of Abnormal and Social Psychology, 1955, 51, 263-268.

Bieri, J. Complexity-simplicity as a personality variable in cognitive and preferential behavior. In D. W. Fiske & S. R. Maddi (Eds.) Functions of varied experience. Homewood, Illinois: The Dorsey Press, Inc., 1961.

Bieri, J., Atkins, A. L., Briar, S., Leaman, R. L., Miller, H., & Tripodi, T. Clinical and social judgment: The discrimination of behavioral information. New York: John Wiley and Sons, 1966.

Blanz, F., & Ghisselli, E. E. The mixed standard scale: A new rating system. Personnel Psychology, 1972, 25, 185-189.

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

Campbell, V. N. Assumed similarity, perceived sociometric balance, and social balance. Unpublished Ph.D. dissertation, University of Colorado, 1960.

Cohen, A. R. Cognitive tuning as a factor affecting impression formation. Journal of Personality, 1961, 29, 235-245.

Craig, G., & Duck, S. W. Similarity, interpersonal attitudes and attraction: The evaluative-descriptive distinction. British Journal of Social and Clinical Psychology, 1977, 16, 15-21.

Crockett, W. H. Cognitive complexity and impression formation. In B. A. Maher (Ed.), Progress in Experimental Personality Research, Vol. 2, New York: Academic Press, 1965.

Estes, S. G. Judging personality from expressive behavior. Journal of Abnormal and Social Psychology, 1938, 33, 217-236.

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

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

Goldstein, K., & Scheerer, M. Abstract and concrete behavior. Psychological Monographs, 1941, 53.

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.

Gordon, M. E. An examination of the relationship between the accuracy and favorability of ratings. Journal of Applied Psychology, 1972, 56, 49-53.

Ivancevich, J. M. Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 1979, 64, 502-508.

Johnston, S., & Centers, R. Cognitive systematization and interpersonal attraction. Journal of Social Psychology, 1973, 90, 95-103.

Kelley, E. L., & Fiske, D. W. The prediction of performance in clinical psychology. Ann Arbor: University of Michigan Press, 1951.

Kelly, G. A. The psychology of personal constructs. Volume I. New York: Norton Press, 1955.

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

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

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

Leventhal, H. Cognitive process and interpersonal predictions. Journal of Abnormal and Social Psychology, 1957, 55, 176-180.

Leventhal, H., & Singer, D. L. Cognitive complexity, impression formation, and impression change. Journal of Personality, 1964, 32, 210-226.

Levy, L. H., & Dugan, R. D. A factorial study of personal constructs. Journal of Consulting Psychology, 1956, 20, 53-57.

Mayo, C. W., & Crockett, W. H. Cognitive complexity and primacy-recency effects in impression formation. Journal of Abnormal and Social Psychology, 1964, 68, 335-338.

Metcalfe, R. J. A. Own and provided constructs in a reptest measure of cognitive complexity. Psychological Reports, 1974, 35, 1305-1306.

Nidorf, L. J., & Crockett, W. H. Cognitive complexity and the integration of conflicting information in written impressions. Journal of Social Psychology, 1965, 66, 165-169.

Olson, J. M., & Partington, J. T. An integrative analysis of two cognitive models of interpersonal effectiveness. British Journal of Social and Clinical Psychology, 1977, 16, 13-14.

Overall, J. E., & Klett, C. J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Plotnick, H. L. The relation between selected personality characteristics of social work students and accuracy in predicting the behavior of clients. Unpublished Ph.D. dissertation, Columbia University, 1961.

Press, A. N., Crockett, W. H., & Rosenkrantz, P. S. Cognitive complexity and the learning of unbalanced social structures. Journal of Personality, 1969, 37, 541-553.

Saal, F. E. Mixed standard rating scale: A consistent system for numerically coding inconsistent response combinations. Journal of Applied Psychology, 1979, 64, 422-428.

Saal, F. E., Downey, R. G., & Lahey, M. A. Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, in press.

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. Measuring cognitive complexity: Developing reliability, validity, and norm tables for a personality instrument. Educational and Psychological Measurement, 1979, 39, 599-612.

Schwab, D. P., Heneman, H. G., III, & DeCotiis, T. A. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Scott, W. A. Cognitive complexity and cognitive flexibility. Sociometry, 1962, 25, 404-414.

Scott, W. A. Cognitive complexity and cognitive balance. Sociometry, 1963, 26, 66-74.

Sechrest, L. B., & Jackson, D. N. Social intelligence and accuracy of interpersonal predictions. Journal of Personality, 1961, 29, 167-181.

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

Smith, S., & Leach, C. A hierarchical measure of cognitive complexity. British Journal of Psychology, 1972, 63, 561-568.

Taft, R. The ability to judge people. In T. L. Whisler & S. F. Harper (Eds.), Performance appraisal: Research and practice.

Tripodi, T., & Bieri, J. Cognitive complexity as a function of own and provided constructs. Psychological Reports, 1963, 13, 26.

Tripodi, T., & Bieri, J. Information transmission in clinical judgements as a function of dimensionality and cognitive complexity. Journal of Personality, 1964, 32, 119-137.

Vannoy, J. S. Generality of cognitive complexity-simplicity as a personality construct. Journal of Personality and Social Psychology, 1965, 2, 385-396.

Warmke, D. L., & Billings, R. S. Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 1979, 64, 124-131.

Zajonc, R. B. The process of cognitive tuning in communication. Journal of Abnormal and Social Psychology, 1960, 60, 159.

APPENDIX I

The Modified Grid Form of the Role Constructs Repertory Test

ILLEGIBLE DOCUMENT

THE FOLLOWING DOCUMENT(S) IS OF POOR LEGIBILITY IN THE ORIGINAL

THIS IS THE BEST COPY AVAILABLE

THIS BOOK CONTAINS NUMEROUS PAGES THAT WERE BOUND WITHOUT PAGE NUMBERS.

THIS IS AS RECEIVED FROM CUSTOMER.

THIS BOOK CONTAINS NUMEROUS PAGES WITH DIAGRAMS THAT ARE CROOKED COMPARED TO THE REST OF THE INFORMATION ON THE PAGE.

THIS IS AS RECEIVED FROM CUSTOMER.

Name_____
Course_____

This exercise is designed to gather information about some of the ways in which you view other people and how you use certain adjectives to describe others.

Please read the instructions carefully and then listen to the experimenter who will explain them. Please try to fill out the form completely and carefully.

INSTRUCTIONS

Attached to this sheet you see a grid of empty squares and labels for the columns and rows of the grid. The labels for the columns represent ten typical roles, or categories of people, that you interact with periodically. Some are members of your family, some are your friends, some you meet at school, etc. For each of these labels, think of the name of a person who best fits that description and write their initials or first name in blank next to the role title. This is only to help you to remember who you are thinking about when you fill out the grid, the actual identity of these people is not required.

The labels for the rows on the grid each consists of two words. These are terms we can use to describe people we know. Each adjective in the set is more or less the opposite of the other. For example, we could describe a person as either outgoing or as shy. You will be asked to describe each of the people on this form with these adjectives. There are also two sets of numbers given at the bottom and the top of these labels. The numbers are L1, L2, L3, R1, R2, R3. These refer to what degree you feel the people you are describing fit or do not fit either of the adjectives in each of the ten pairs of adjectives. For example, if you feel the person that you are thinking about for the role of friend is very shy, you would mark an R3 in the box corresponding to the column with the label "friend" and the row labeled "out-going -- shy". If the person you were thinking of for that role were only slightly shy, you would mark an R1 in that box, and so on.

You are asked to describe how you feel about each of the ten persons listed here on each of the ten pairs of adjectives. Thus, you can take the first role, and mark your feelings on each the rows, go on to the second role and mark your feelings on each of rows, etc until the grid is completed. Use the same person for the role throughout the entire grid. If you are unsure about any of the persons you have put in the roles and don't know the person well enough to have a definite opinion, put down the number and letter that you feel most appropriately describes the person based on the knowledge you do have. Remember there are no right and wrong answers and that this sheet will be used solely for research purposes.

Thank you for your cooperation, If there are any questions please
ASK THE EXPERIMENTER

1. Yourself (____)	2. Person you dislike (____)	3. Mother (____)	4. Person you'd like to help (____)	5. Father (____)	6. Friend of same sex (____)	7. Friend of opposite sex or spouse (____)	8. Person with whom you feel most uncomfortable (____)	9. Boss (____)	10. Person difficult to understand (____)

	L3	L2	L1	--	R1	R2	R3
1.	Outgoing				Shy		
2.	Maladjusted				Adjusted		
3.	Decisive				Indecisive		
4.	Excitable				Calm		
5.	Interested in others				Self-absorbed		
6.	Ill-humored				Cheerful		
7.	Responsible				Irresponsible		
8.	Inconsiderate				Considerate		
9.	Independent				Dependent		
10	Dull				Interesting		
	L3	L2	L1	--	R1	R2	R3

CONFIDENTIAL
For Research Purposes Only

APPENDIX II

The Role Constructs Repertory Test

This exercise is designed to gather information about some of the ways in which you view other people and how you use adjectives to describe others.

Please read the instructions carefully and then listen to the experimenter who will explain them. Please try to fill out this form completely and carefully.

INSTRUCTIONS

Attached to this sheet you will find a grid of empty squares and labels for the columns of the grid. The labels for the columns represent 15 typical roles, or categories of people, that you interact with. Some are members of your family some are your friends, some you meet at school, etc. For each of these labels, think of the name of the person who best fits that description and write their initials or first name in the blank next to that role title. This is only to help you remember who you are thinking about as you fill out the grid, the actual identity of the persons is not required.

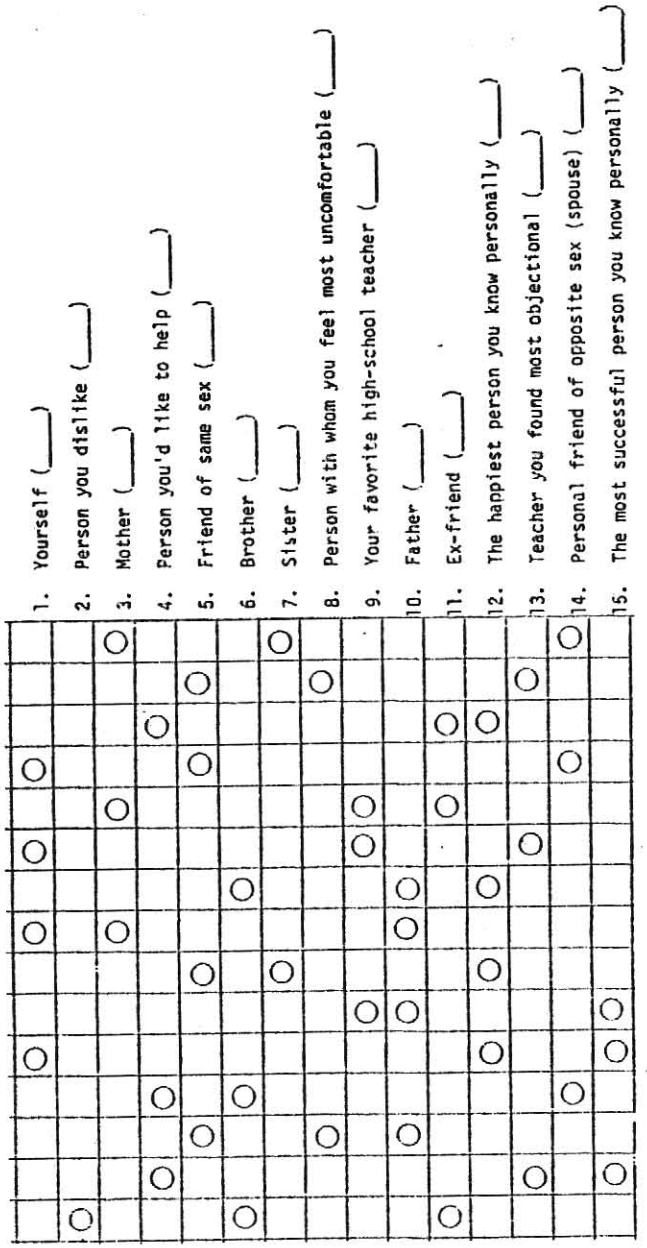
In each of the rows of the grid you will find three circles. The sets of three circles in each row indicate the roles which you are to consider together. For each row, consider the three persons whom you have listed in the spaces above. Determine what important ways two of these people are alike and, at the same time, essentially different from the third. When making your decisions about important aspects please avoid using information which is descriptive of the persons' appearance. For example, try to avoid the use of labels such as male/female, blonde/brunnette, blue-eyed/brown-eyed, etc.

After you have decided what important way two of these people ar alike and yet, different from the third write the label for that aspect in the blank which is labled "construct". Indicate which of the two people share that construct by placing a check (✓) in the circle beneath their name. Then write down the label which you believe to be the opposite of the construct in the balnk marked as "opposite".

On the following page please transfer the persons names or initials and the constructs/opposites into the appropriate spaces being sure to match the grid on the previous page. Then, in each of the boxes provided on this grid rate each of the persons you've listed on all of the constructs. That is, give each person a numerical rating on all of the constructs you named. On this scale a rating of 1 indicates that the construct is highly descriptive of the individual being rated and a rating of 5 indicates that the opposite is descriptive of them. A 3 will indicate that the person is somewhere between the two.

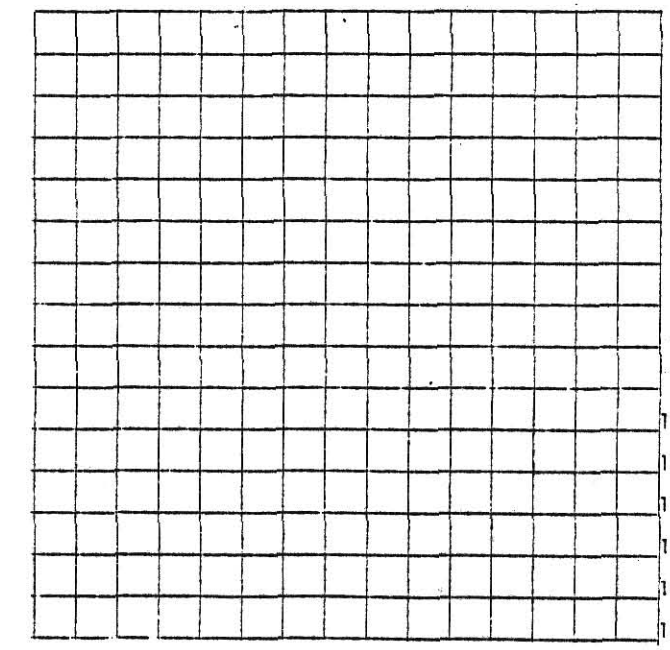
You are asked to describe how you feel about each of the 15 persons listed on each of the 15 constructs/opposites. Thus, you can take the first role,and mark your feelings on each of the rows, go on to the second role and mark your feelings on each of the rows, etc. until the entire grid is completed. Use the same person for the role throughout the entire grid. If you are unsure about any of the persons you have put in a role or don't know the person well enough to have a definite opinion, put down the number which you feel most appropriately describes the person based on the knowledge you do have. Remember there are no right and wrong answers and that this sheet will be used solely for research purposes.

Thank you for your cooperation; If there are any questions
ASK THE EXPERIMENTER

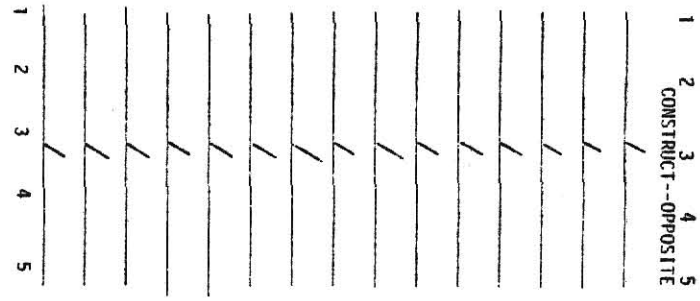


CONSTRUCT--OPPOSITE

_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____
_____/_____



1. Yourself (_____)
2. Person you dislike (_____)
3. Mother (_____)
4. Person you'd like to help (_____)
5. Friend of same sex (_____)
6. Brother (_____)
7. Sister (_____)
8. Person with whom you feel most uncomfortable (_____)
9. Your favorite high-school teacher (_____)
10. Father (_____)
11. Ex-friend (_____)
12. The happiest person you know personally (_____)
13. Teacher you found most objectional (_____)
14. Personal friend of opposite sex (spouse) (_____)
15. The most successful person you know personally (_____)



APPENDIX III

Behaviorally Anchored Rating Scale

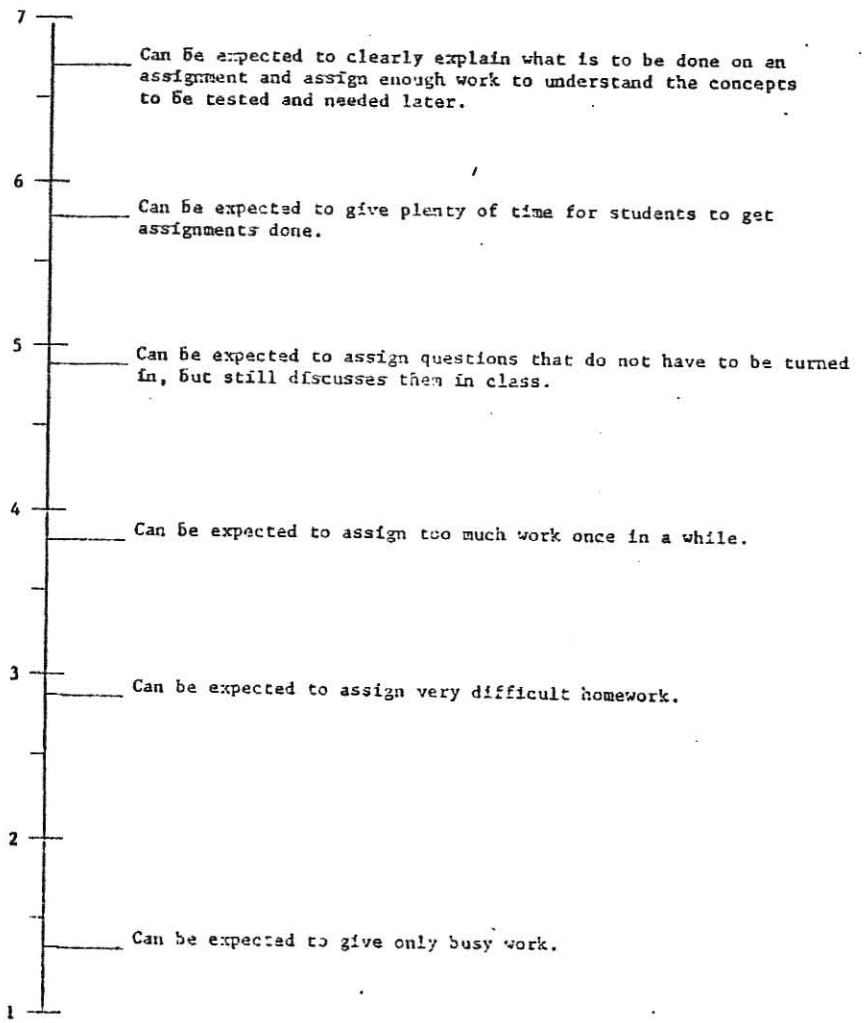
The following pages include nine categories concerning various aspects of teaching with definitions provided for each one. The categories and definitions resulted from meetings held with groups of college students. Below each category is a seven-point rating scale with behavioral statements located at various point along the scale. All of the statements were written by students and each statements location on the scale was determined by student's evaluation of the level of performance it best represented.

Please read the definition for each category, carefully, and the behavioral statement shown on the right of the accompanying scales. Then compare those behaviors of your instructor which apply to each particular category against the behaviors on the scales. Finally, use these behaviors as references or aids in helping you to determine the rating that you feel your instructor deserves. That is, based upon the behaviors that you have seen, rate your instructor according to the type of behavior that you would expect of him/her when compared to the behavior on the scale. It is possible that your instructor has never exhibited any of the behaviors shown on the scale. However, based upon those behaviors that you have seen exhibited by your instructor and which are applicable to the category your rating, rate him/her on the level of performance you would expect from him/her relative to the statements on the scales.

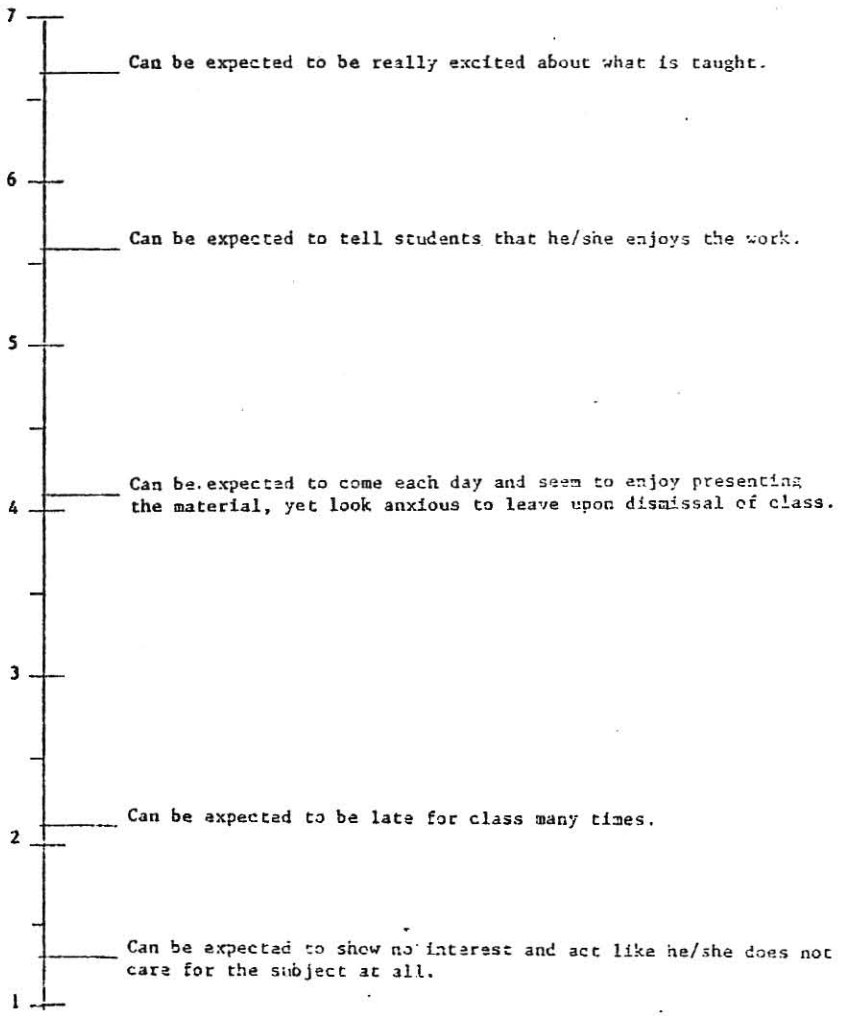
Please make a single mark anywhere on the vertical line. It is not necessary to make a mark only where a statement is located. The statements are merely references against which your are to compare the actual obsereved behavior of the instructor.

At the end of the set of scales there is one additional five-point scale which you should fill out. After rating your instructor on all the categories decide how sure you are of these ratings and circle the one number which best describes how sure you are of all the ratings. In other words how certain are you of the total set of ratings that you have given.

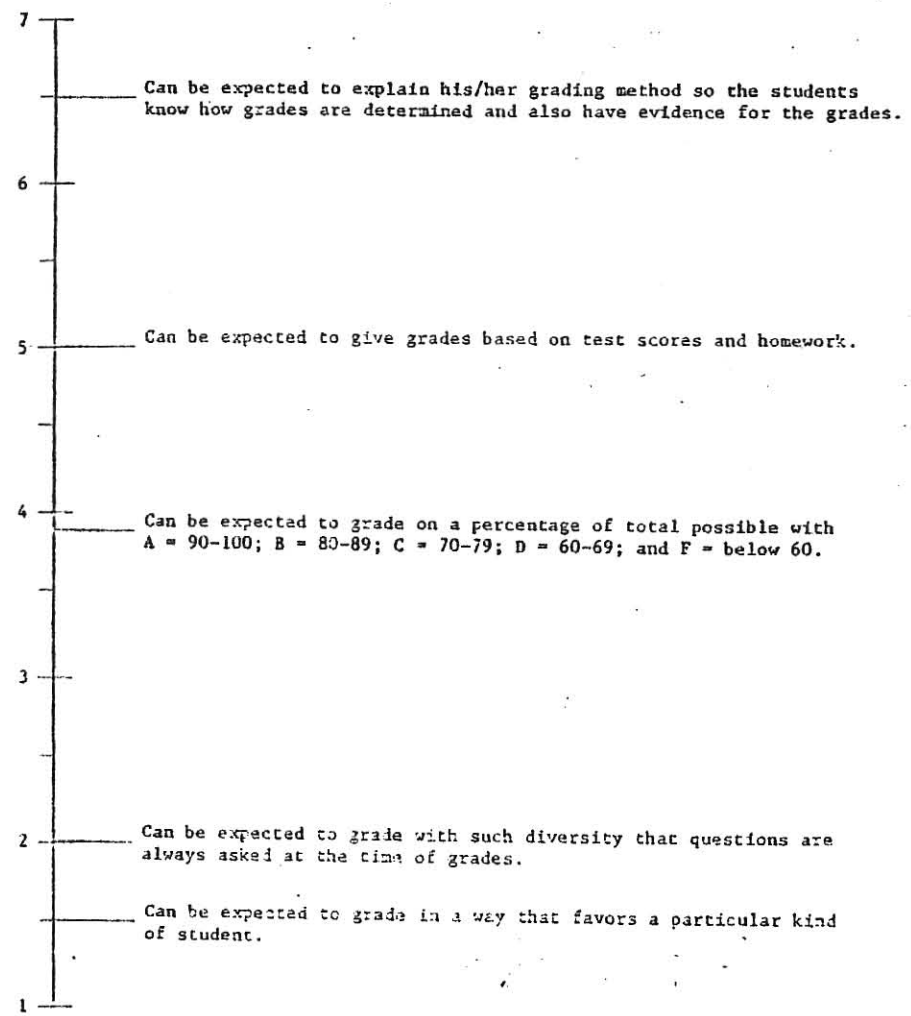
ASSIGNMENTS: Extent to which the instructor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.



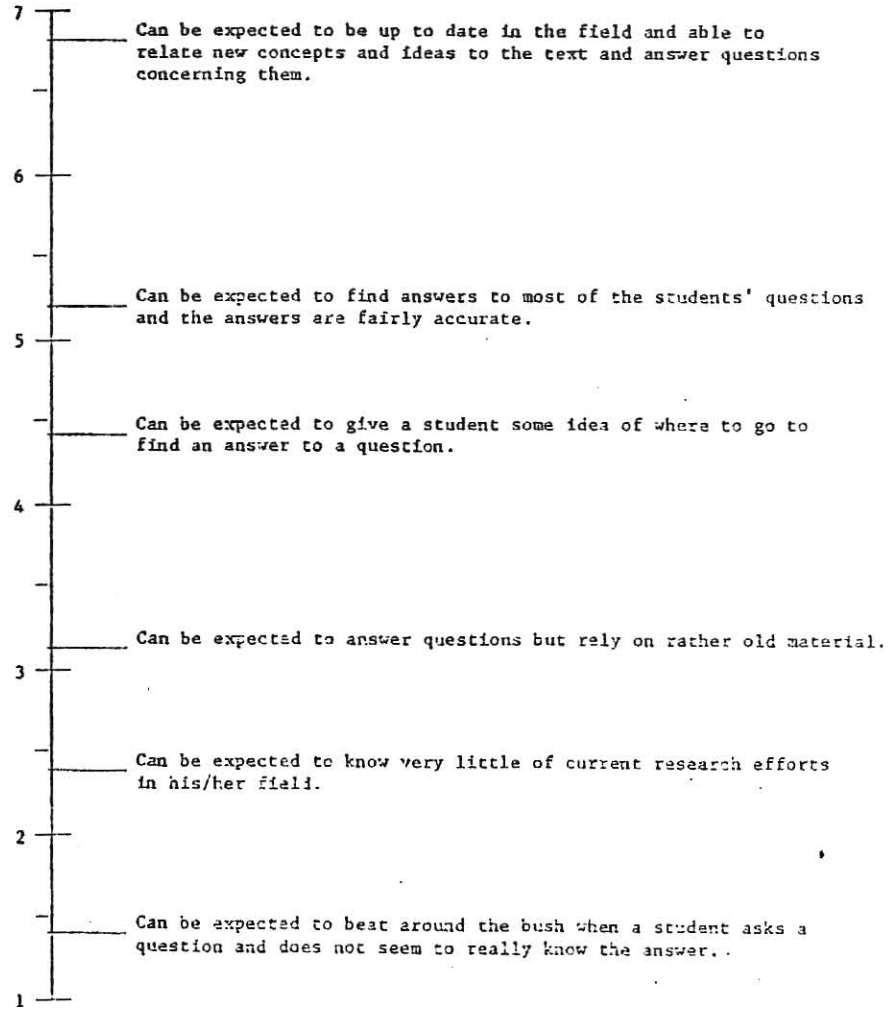
ATTITUDE TOWARDS SUBJECT: Extent to which the instructor shows personal interest in the material and displays a positive attitude towards teaching the subject.



GRADES: Extent to which the instructor's grading practices remain consistent and free of confusion and are also fair.



INSTRUCTOR KNOWLEDGE: Extent to which the instructor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.

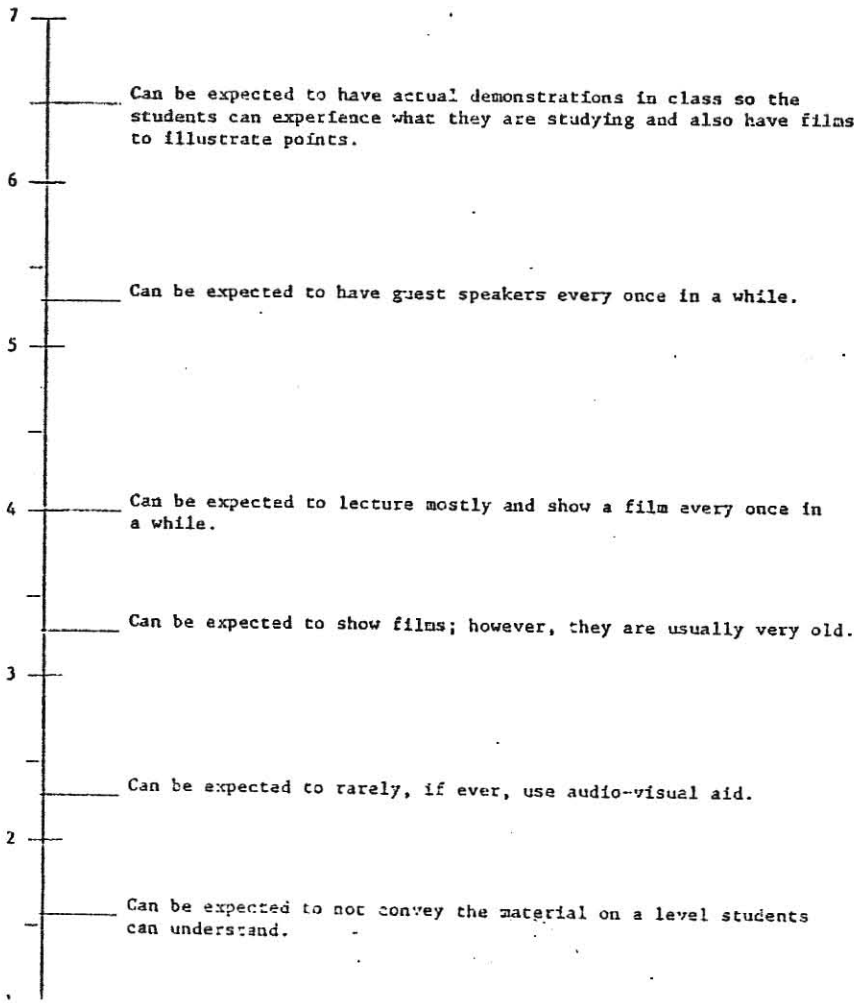


ILLEGIBLE

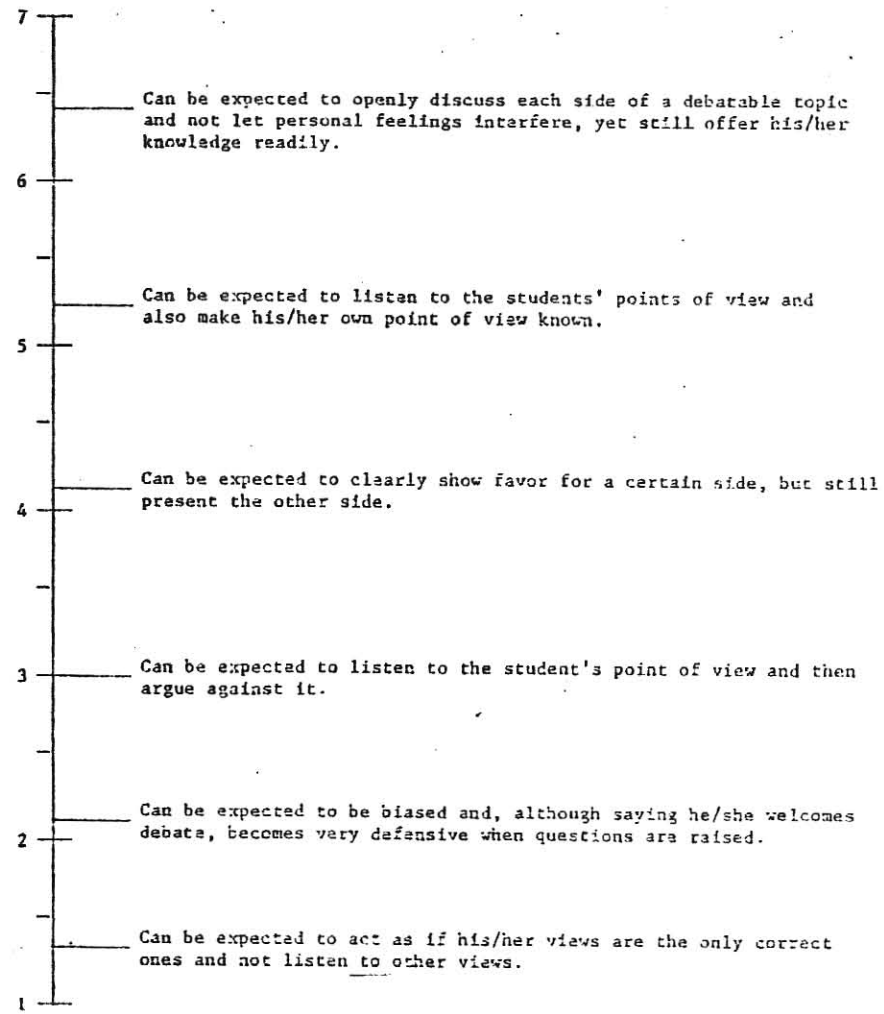
THE FOLLOWING DOCUMENT (S) IS ILLEGIBLE DUE TO THE PRINTING ON THE ORIGINAL BEING CUT OFF

ILLEGIBLE

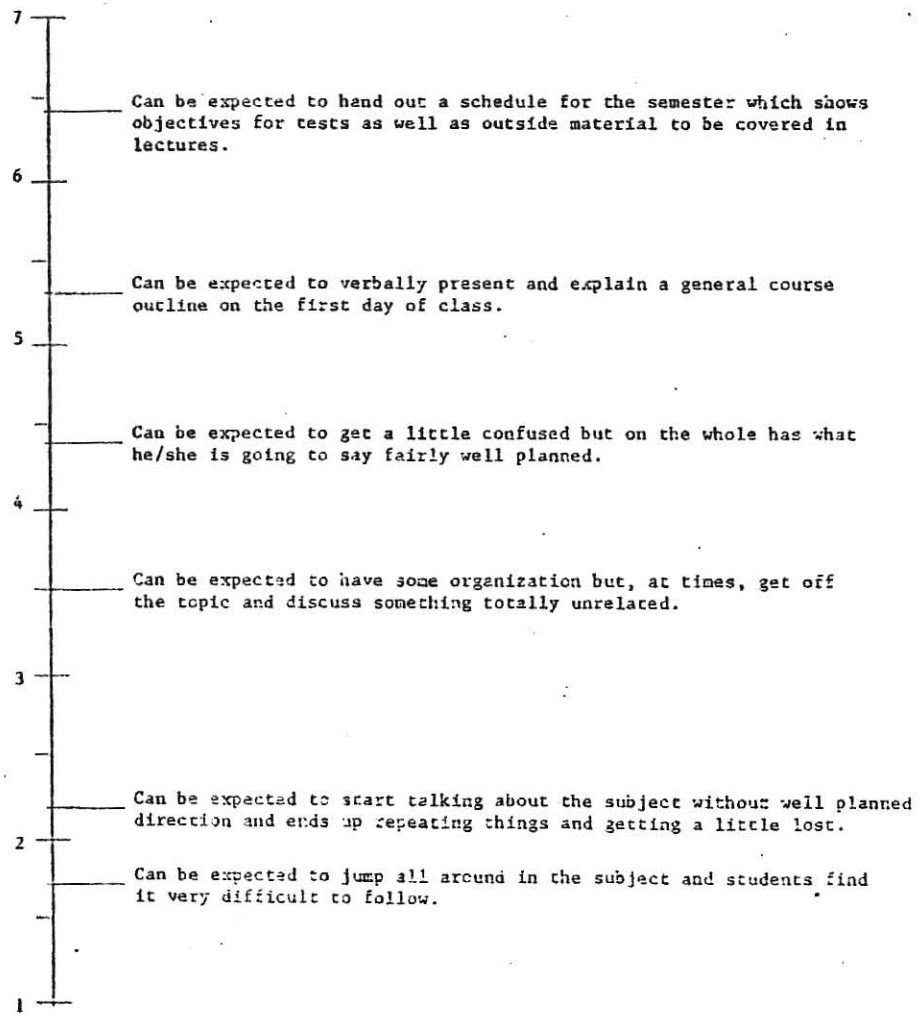
MANNER OF PRESENTATION: Extent to which the instructor's methods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level students can understand.



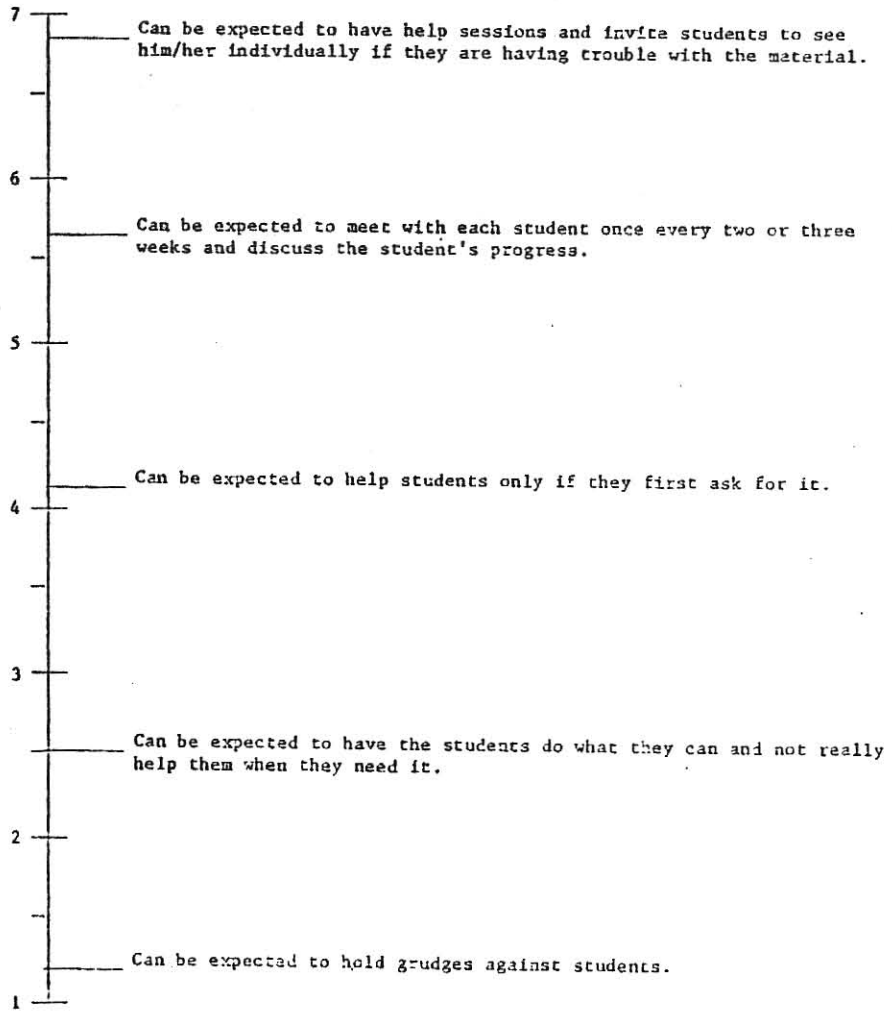
OBJECTIVENESS: Extent to which the instructor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.



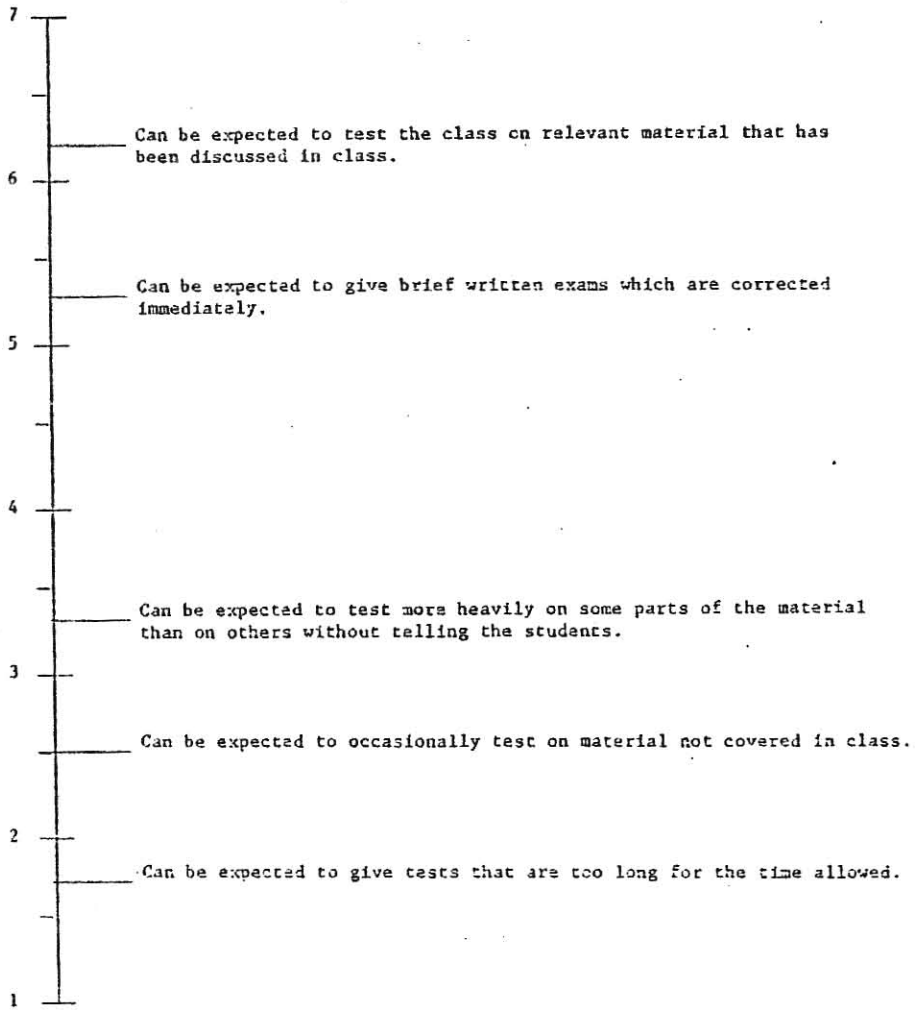
ORGANIZATION: Extent to which the instructor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.



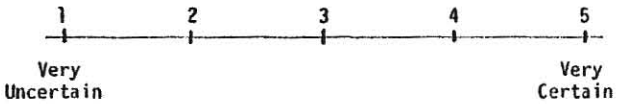
STUDENT-TEACHER RELATIONS: Extent to which the instructor shows a true, sincere concern for the welfare of the students through such things as dependability, availability for help, and consideration of student feelings; establishing rapport with the students.



TESTS: Extent to which the instructor writes clear, unambiguous questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.



Certainty of Ratings: Based on the total set of ratings that you have just given your instructor, how certain are you that they adequately reprsent his/her level of teaching performance this semester?



APPENDIX IV

Mixed Standard Rating Scale

Listed on the following pages are a number of descriptions of behaviors concerning various aspects of teaching. To the right of each behavioral statement are three boxes. The box on the far left corresponds to "The instructor is worse than this statement"; the center box corresponds to "This statement fits the instructor"; and the box on the far right corresponds to "The instructor is better than this statement". Carefully read and consider each statement and determine in your own mind the answer to the following question: Is the instructor who I am rating "better than" this statement, "worse than" this statement, or does this statement "fit" the instructor?

If you believe the instructor's behavior is better than the statement you have just read, mark an X in the box on the far righ which corresponds with "The instructor is better than this statement". If you believe that the instructor's behavior is worse than that described in the statement you have just read, mark an X in the box on the left which corresponds with "The instructor is worse than this statement. If you believe that the statement you have just read adequately describes your instructor, mark an X in the center box which corresponds with "This statement fits the instructor". For each of the behavioral statements on the following pages mark only <u>one</u> box. Also, be sure to make a mark along side of every behavioral statement.

At the end of the set of statements there is an additional five-point scale which you should fill out. After rating your instructor on all of the behavior statements decide how sure you are of the ratings you made and make <u>one</u> mark along the line which best describes how sure you are of all the ratings. In other words, how certain are you of the total set of ratings that you have given?

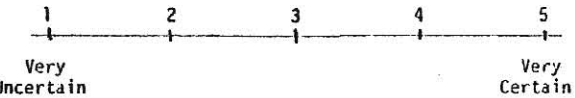
Statements	Ratings: The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
1. Can be expected to give assignments that offer at least some understanding of the subject matter.			
2. Can be expected to give a vague sheet of chapter order and then rarely cover the material on schedule or give tests on projected days.			
3. Can be expected to use interesting films and discuss them afterwards in order to clarify major points.			
4. Can be expected to be completely close-minded and students that openly disagree with him/her have a difficult time in the class.			
5. Can be expected to present material that sometimes overlaps with material that has already been presented.			
6. Can be expected to invite students to come and talk with him/her after class or during office hours whenever they have a problem with the class.			
7. Can be expected to show an interest in the subject, but seem to lack any great involvement.			
8. Can be expected to curve the grades according to the difficulty of the test and the overall student scores.			
9. Can be expected to not let students know how grades are derived.			
10. Can be expected to seem not to know what he/she is talking about and can't answer questions.			

Statements	Ratings: The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
11. Can be expected to give assignments that relate to the subject matter and provide further understanding of it.			
12. Can be expected to test in order to see if students have read the text and kept up with the assignments.			
13. Can be expected to add current material to class presentations for use as new examples.			
14. Can be expected to go through the material very fast and allow no time for questions.			
15. Can be expected to leave the lecture material only occasionally to give an interesting sideview.			
16. Can be expected to show the various points of view, allow discussion, not put down anyone, listen, and let students decide for themselves.			
17. Can be expected to not even care if students do not understand the material.			
18. Can be expected to show that he/she enjoys the subject and job and students can tell.			
19. Can be expected to have a student come to his/her office after class or wait until the next class for an answer to a question.			
20. Can be expected to vaguely give the student an idea of what is to be done on the next assignment and not really say when it's due.			

Ratings

Statements	The instructor is worse than this statement.	This statement fits the instructor.	The instructor is better than this statement.
21. Can be expected to discuss each side of a controversial topic but lean towards his/her beliefs and try to convince others of them..			
22. Can be expected to use only test scores for determining grades.			
23. Can be expected to show dislike for the subject and teaching and act as though he/she just got stuck with it.			
24. Can be expected to allow the students ample time to answer test questions carefully.			
25. Can be expected to have an outlined, detailed schedule of each day's activities to present to the students at the beginning, but it can be changed if necessary.			
26. Can be expected to inform students of fairly recent developments in the area during the past few years, but is not up to date on present studies.			
27. Can be expected to have a lot of test questions pertaining to subject matter that was barely even mentioned in class or not mentioned at all.			

Certainty of Ratings: Based on the total set of ratings that you have just given your instructor, how certain are you that they adequately reprsent his/her level of teaching performance this semester?



APPENDIX V

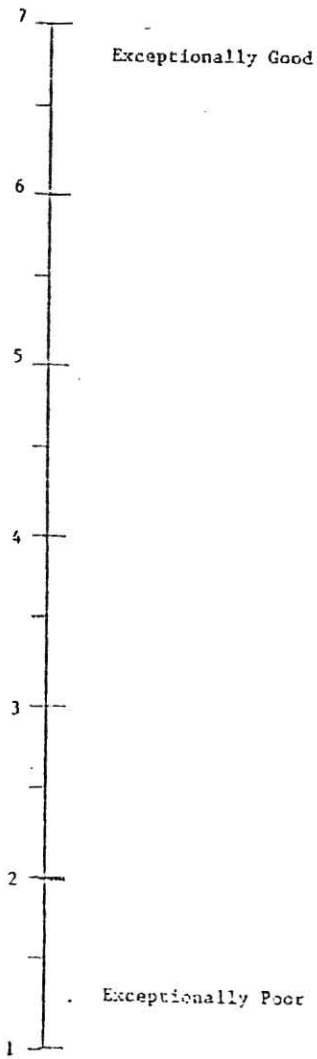
Graphic Rating Scale

The following pages contain nine categories concerning various aspects of teaching with definitions provided for each one. The categories and definitions resulted from meetings held with groups of college students. Below each category is a seven-point rating scale.

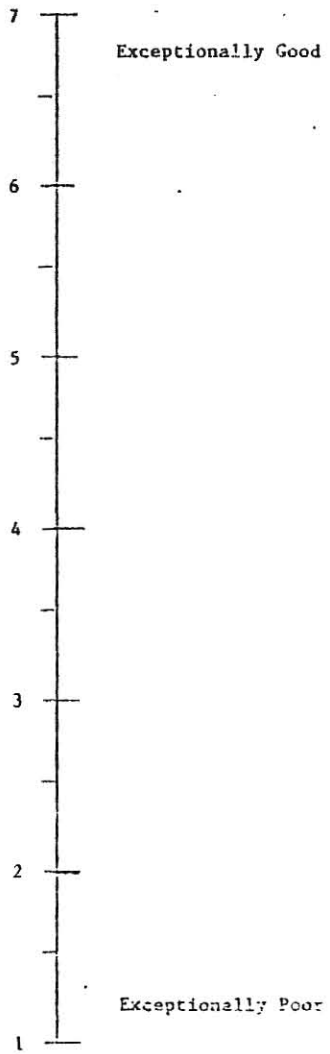
Please read the definition for each category carefully. Considering only those behaviors of your instructor which apply to each particular category, determine the rating that you feel your instructor deserves. Then make a single mark anywhere along the vertical line which you feel reflects your assessment of your instructor's performance, follow this procedure for every category.

At the end of the set of scales there is one additional five-point scale which you should fill out. After rating your instructor on all of the categories, decide how sure you are of these ratings and make one mark along the line which best describes how sure you are of these ratings. In other words, how certain are you of the total set of ratings that you have just given?

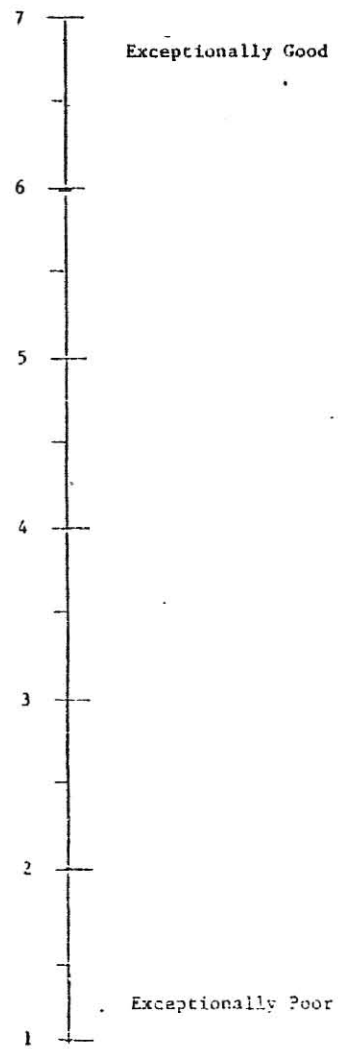
ASSIGNMENTS: Extent to which the instructor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.



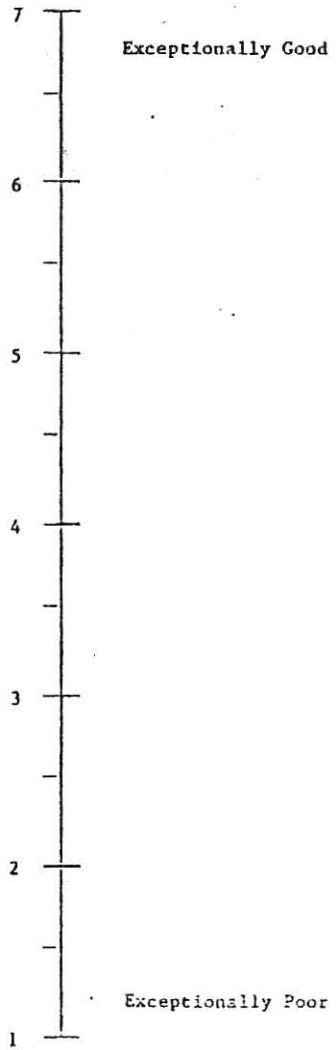
ATTITUDE TOWARDS SUBJECT: Extent to which the instructor shows personal interest in the material and displays a positive attitude towards teaching the subject.



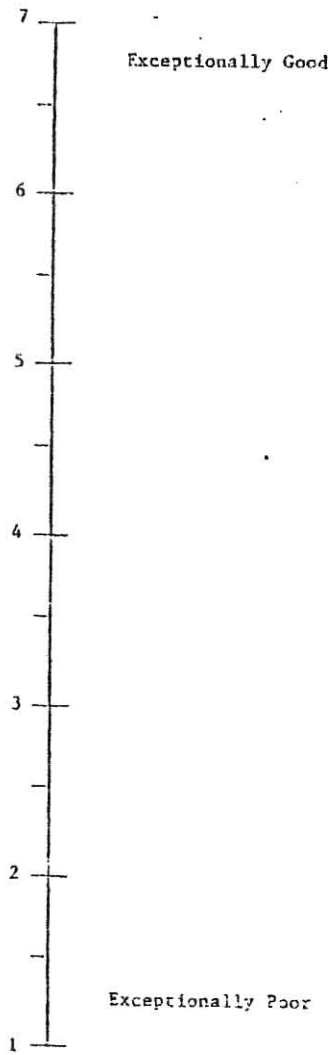
GRADES: Extent to which the instructor's grading practices remain consistent and free of confusion and are also fair.



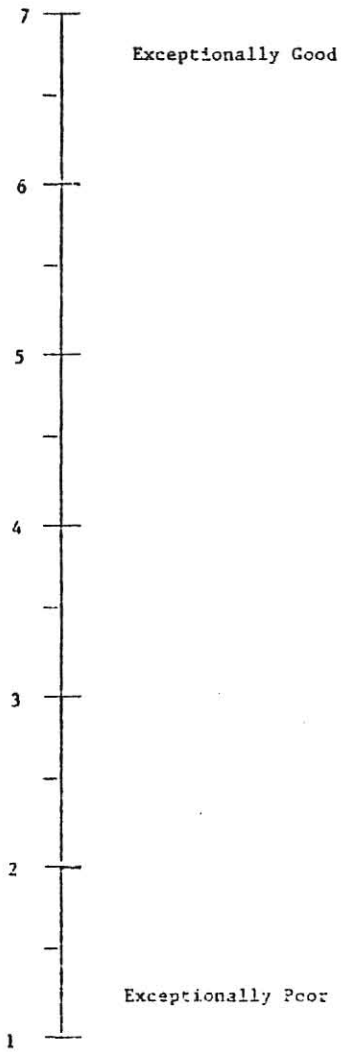
INSTRUCTOR KNOWLEDGE: Extent to which the instructor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.



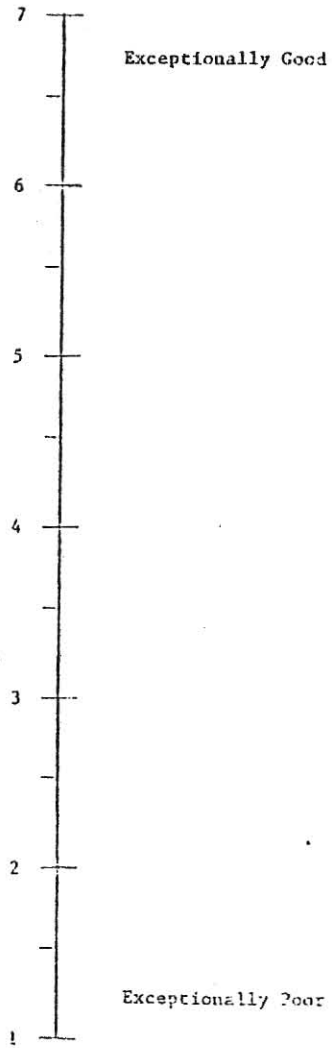
MANNER OF PRESENTATION: Extent to which the instructor's methods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level students can understand.



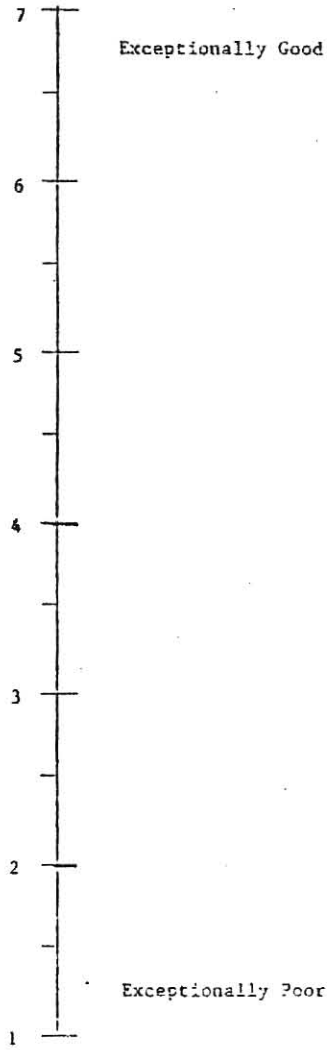
OBJECTIVENESS: Extent to which the instructor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.



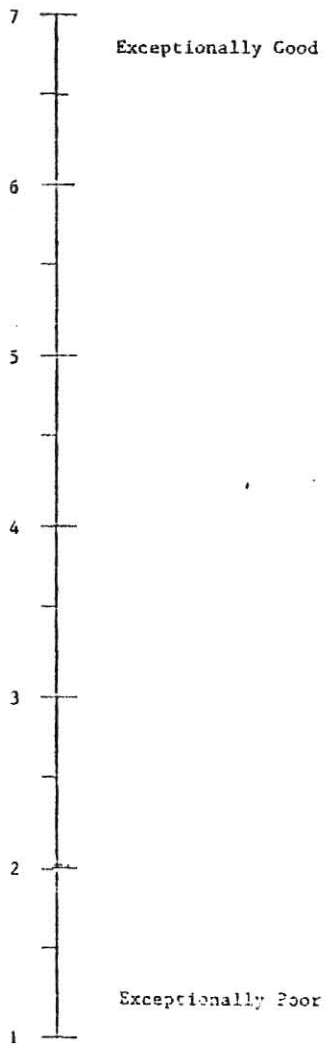
ORGANIZATION: Extent to which the instructor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.



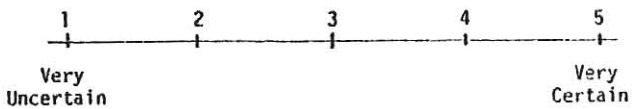
STUDENT-TEACHER RELATIONS: Extent to which the instructor shows a true, sincere concern for the welfare of the students through such things as dependability; availability for help, and consideration of student feelings; establishing rapport with the students.



TESTS: Extent to which the instructor writes clear, unambiguous questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.



Certainty of Ratings: Based on the total set of ratings that you have just given your instructor, how certain are you that they adequately reprsent his/her level of teaching performance this semester?



APPENDIX VI

Alternate Rating Scale

The following pages contain nine categories concerning various aspects of teaching with definitions provided for each one. The categories and definitions resulted from meetings held with groups of college students. Below each category there is a three-point rating.

Please read the definition for each category carefully. Considering only those behaviors of your instructor which apply to each particular category, determine the rating that you feel your instructor deserves. Then mark an X on the line which you feel best reflects your assessment of your instructor's performance, follow this procedure for every category.

At the end of the set of scales there is an additional five-point scale which you should fill out. After rating your instructor on all of the categories, decide how sure you are of these ratings and make one mark along the line which best describes how sure you are of these ratings. In other words, how certain are you of the total set of ratings you have just given?

Assignments: Extent to whivh the instructor is clear on what is to be done, avoids assigning excessive amounts, and provides assignments which contribute to the understanding of the subject matter rather than just providing busy work.

_____ Above Average

_____ Average

_____ Below Average

Attitude Towards Subject: Extent to which the instructor shows personal interest in the material and displays a positive attitude towards teaching the subject.

_____ Above Average

_____ Average

_____ Below Average

Grades: Extent to which the instructor's grading practices remain consistent and free of confusion and are also fair.

_____ Above Average

_____ Average

_____ Below Average

Instructor Knowledge: Extent to which the instructor is aware of current material related to the course or to his/her field and is able to accurately answer or direct the student to specific sources that will answer questions concerning the subject matter.

_____ Above Average

_____ Average

_____ Below Average

Manner of Presentation: Extent to which the instructor's mehtods of presentation and use of audio-visual aids help emphasize and clarify important points; ability to present material clearly and concisely on a level students can understand.

_____ Above Average

_____ Average

_____ Below Average

Objectiveness: Extent to which the instructor remains objective and presents a fair treatment of all points of view on controversial or debatable topics.

_____ Above Average

_____ Average

_____ Below Average

Organization: Extent to which the instructor arranges the subject matter and course objectives in an orderly and logical sequence for thorough coverage.

_____ Above Average

_____ Average

_____ Below Average

Student-Teacher Relations: Extent to which the instructor shows a true, sincere concern for the welfare fo the students through such things as dependability, availability for help, and consideration of students feelings; establishing rapport with the students.

_____ Above Average

_____ Average

_____ Below Average

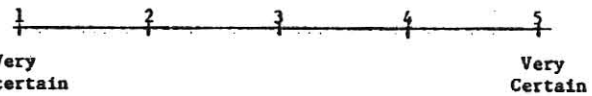
Tests: Extent to which the instructor writes clear, unambiguous test questions that relate to and are representative of in-class material and outside readings which were stressed adequately in class.

_____ Above Average

_____ Average

_____ Below Average

Certainty of Ratings: Based on the total set of ratings that you have just given your instructor, how certain are you that they adequately represent his/her level of teaching performance this semester?



APPENDIX VII

Source Tables for Analyses of Variance
Modified Grid Form of REP Test

Source Tables for Leniency: Modified Grid Form of REP Test

Complexity Main Effects

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.028	1	.028	≤ 1	n.s.
2	.011	1	.011	<1	n.s.
3	.059	1	.059	<1	n.s.
4	.654	1	.654	1.96	n.s.
5	.584	1	.584	1.27	n.s.
6	1.419	1	1.419	2.98	n.s.
7	.169	1	.169	<1	n.s.
8	.171	1	.171	<1	n.s.
9	.154	1	.154	<1	n.s.
Mean	.050	1	.050	<1	n.s.

Complexity x Scale Format Interactions

Dimension	Sums of Squares	DF	Mean Square	F	p
1	1.036	3	.345	2.83	n.s.
2	.308	3	.103	1.09	n.s.
3	.479	3	.159	<1	n.s.
4	.229	3	.076	<1	n.s.
5	.453	3	.151	<1	n.s.
6	.439	3	.146	<1	n.s.
7	.215	3	.072	<1	n.s.
8	.082	3	.027	<1	n.s.
9	.061	3	.020	<1	n.s.
Mean	.053	3	.018	1.30	n.s.

Source Tables for Restriction of Range: Modified Grid Form of REP Test

Complexity Main Effects

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.415	1	.415	2.97	n.s.
2	.591	1	.591	3.88	n.s.
3	.011	1	.011	<1	n.s.
4	.813	1	.813	5.07	<.05
5	.081	1	.081	<1	n.s.
6	.403	1	.403	2.53	n.s.
7	.506	1	.506	3.09	n.s.
8	.518	1	.518	3.07	n.s.
9	.088	1	.088	<1	n.s.
Mean	.291	1	.291	4.08	<.05

Complexity x Scale Format Interactions

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.067	3	.022	<1	n.s.
2	.060	3	.020	<1	n.s.
3	.013	3	.013	<1	n.s.
4	.042	3	.014	<1	n.s.
5	.305	3	.102	1.56	n.s.
6	.221	3	.074	1.51	n.s.
7	.081	3	.027	<1	n.s.
8	.018	3	.006	<1	n.s.
9	.064	3	.021	<1	n.s.
Mean	.018	3	.006	<1	n.s.

Source Table for Halo: Modified Grid Form of REP Test

Source	Sums of Squares	DF	Mean Square	F	p
Complexity	.055	1	.055	1.07	n.s.
Complexity x Scale Format	.003	3	.001	< 1	n.s.

APPENDIX VIII

Source Tables for Analyses of Variance
Factor Analysis of REP Test

Source Tables for Leniency: Factor Analysis of REP Test

Complexity Main Effects

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.0003	1	.0003	≤ 1	n.s.
2	.334	1	.334	< 1	n.s.
3	.298	1	.298	< 1	n.s.
4	.129	1	.129	< 1	n.s.
5	.594	1	.594	1.30	n.s.
6	.088	1	.088	< 1	n.s.
7	.060	1	.060	< 1	n.s.
8	.005	1	.005	< 1	n.s.
9	.016	1	.016	< 1	n.s.
Mean	.002	1	.002	< 1	n.s.

Complexity x Scale Format Interactions

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.636	3	.212	<1	n.s.
2	.068	3	.023	≤ 1	n.s.
3	.061	3	.020	<1	n.s.
4	.448	3	.150	<1	n.s.
5	.043	3	.014	<1	n.s.
6	.289	3	.096	<1	n.s.
7	.327	3	.109	<1	n.s.
8	.565	3	.188	<1	n.s.
9	.133	3	.044	<1	n.s.
Mean	.003	3	.001	<1	n.s.

Source Tables for Restriction of Range: Factor Analysis of REP Test

Complexity Main Effects

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.206	1	.206	1.45	n.s.
2	.065	1	.065	< 1	n.s.
3	.153	1	.153	1.05	n.s.
4	.073	1	.073	< 1	n.s.
5	.005	1	.005	< 1	n.s.
6	.556	1	.556	3.53	n.s.
7	.029	1	.029	< 1	n.s.
8	.754	1	.754	4.53	< .05
9	.239	1	.239	1.38	n.s.
Mean	.157	1	.157	2.16	n.s.

Complexity x Scale Format Interactions

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.223	3	.074	1.68	n.s.
2	.026	3	.009	< 1	n.s.
3	.030	3	.010	< 1	n.s.
4	.083	3	.028	< 1	n.s.
5	.306	3	.102	< 1	n.s.
6	.099	3	.033	< 1	n.s.
7	.116	3	.052	1.13	n.s.
8	.048	3	.016	< 1	n.s.
9	.128	3	.043	< 1	n.s.
Mean	.012	3	.004	< 1	n.s.

Source Table for Halo: Factor Analysis of REP Test

Source	Sums of Squares	DF	Mean Square	F	p
Complexity	.0003	1	.0003	<1	n.s.
Complexity x Scale Format	.046	3	.015	1.05	n.s.

APPENDIX IX

Source Tables for Analyses of Variance
Sorting Task

Source Tables for Leniency: Sorting Task

Complexity Main Effects

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.797	1	.797	2.89	n.s.
2	.633	1	.633	1.77	n.s.
3	.314	1	.314	< 1	n.s.
4	.052	1	.052	< 1	n.s.
5	.006	1	.006	< 1	n.s
6	.232	1	.232	< 1	n.s.
7	.004	1	.004	< 1	n.s.
8	.032	1	.032	< 1	n.s.
9	.0002	1	.0002	< 1	n.s.
Mean	.027	1	.027	< 1	n.s.

Complexity x Scale Format Interactions

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.259	3	.086	<1	n.s.
2	.084	3	.028	<1	n.s.
3	.476	3	.158	1.00	n.s.
4	.277	3	.094	<1	n.s.
5	.472	3	.157	<1	n.s.
6	.853	3	.284	2.80	n.s.
7	.352	3	.117	<1	n.s.
8	1.706	3	.568	3.79	n.s.
9	.525	3	.175	1.07	n.s.
Mean	.013	3	.004	<1	n.s.

Source Tables for Restriction of Range: Sorting Task

Complexity Main Effects

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.061	1	.061	< 1	n.s.
2	.001	1	.001	< 1	n.s.
3	.049	1	.049	< 1	n.s.
4	.061	1	.061	< 1	n.s.
5	.007	1	.007	< 1	n.s.
6	.216	1	.216	1.34	n.s.
7	.089	1	.089	< 1	n.s.
8	.509	1	.509	3.02	n.s.
9	.170	1	.170	1.00	n.s.
Mean	.076	1	.076	1.04	n.s.

Complexity x Scale Format Interactions

Dimension	Sums of Squares	DF	Mean Square	F	p
1	.071	3	.024	<1	n.s.
2	.027	3	.009	<1	n.s.
3	.089	3	.030	<1	n.s.
4	.043	3	.014	<1	n.s.
5	.364	3	.121	1.87	n.s.
6	.178	3	.059	1.21	n.s.
7	.127	3	.042	<1	n.s.
8	.078	3	.026	<1	n.s.
9	.206	3	.069	1.17	n.s.
Mean	.057	3	.019	1.91	n.s.

Source Table for Halo: Sorting Task

Source	Sums of Squares	DF	Mean Square	F	p
Complexity	.058	1	.058	1.07	n.s.
Complexity x Scale Format	.055	3	.018	1.11	n.s.

APPENDIX X

Source Tables for Main Effects of Rating Scale and Main Effects of Ratees on Each Complexity Measure

Source Tables for Main Effect of Rating Scale Format:
Modified Grid Form of REP Test

Leniency Analyses[a]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	6.439	3	2.279	18.68	<.01
2	1.942	3	.647	6.91	<.01
3	4.712	3	1.571	9.91	<.01
4	2.425	3	.809	8.21	<.01
5	3.981	3	1.327	7.02	<.01
6	1.072	3	.357	2.72	<.05
7	10.071	3	3.357	22.38	<.01
8	2.726	3	.909	5.84	<.01
9	6.108	3	2.036	12.37	<.01
Mean	.745	3	.248	18.14	<.01

Halo Analysis[b]

Source	Sums of Squares	DF	Mean Square	F	p
Scale Format	7.904	3	2.635	160.63	<.01

Range Restriction Analyses[b]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	8.01	3	2.67	59.64	<.01
2	3.67	3	1.22	26.73	<.01
3	4.14	3	1.38	26.23	<.01
4	5.97	3	1.99	50.26	<.01
5	4.77	3	1.59	24.42	<.01
6	3.61	3	1.20	24.60	<.01
7	4.86	3	1.62	35.01	<.01
8	5.34	3	1.78	36.12	<.01
9	4.18	3	1.39	23.51	<.01
Mean	4.53	3	1.51	150.16	<.01

[a] AS < MSS < GRS = BARS

[b] AS > MSS = GRS = BARS

Source Tables for Main Effect of Rating Scale Format:
Factor Analysis of REP Test

Leniency Analyses[a]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	6.439	3	2.279	17.29	< .01
2	1.942	3	.647	6.91	< .01
3	4.712	3	1.571	9.24	< .01
4	2.425	3	.809	7.49	< .01
5	3.981	3	1.327	6.83	< .01
6	1.072	3	.357	3.69	< .01
7	10.071	3	3.357	22.38	< .01
8	2.726	3	.909	5.84	< .01
9	6.108	3	2.036	11.39	< .01
Mean	.745	3	.248	17.91	< .01

Halo Analysis[b]

Source	Sums of Squares	DF	Mean Square	F	p
Scale Format	7.900	3	2.635	160.63	< .01

Range Restriction Analyses[b]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	8.01	3	2.67	60.39	< .01
2	3.67	3	1.22	26.66	< .01
3	4.14	3	1.38	26.28	< .01
4	5.97	3	1.99	50.26	< .01
5	4.77	3	1.59	24.32	< .01
6	3.61	3	1.20	24.38	< .01
7	4.86	3	1.62	35.29	< .01
8	5.34	3	1.78	36.12	< .01
9	4.18	3	1.39	23.60	< .01
Mean	4.53	3	1.51	149.83	< .01

[a] AS < MSS < GRS = BARS

[b] AS > MSS = GRS = BARS

Source Tables for Main Effect of Rating Scale Format:
Sorting Task

Leniency Analyses[a]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	6.439	3	2.279	18.73	< .01
2	1.942	3	.647	6.85	< .01
3	4.712	3	1.571	9.88	< .01
4	2.425	3	.809	8.24	< .01
5	3.981	3	1.327	7.08	< .01
6	1.072	3	.357	2.81	< .01
7	10.071	3	3.357	22.21	< .01
8	2.726	3	.909	5.84	< .01
9	6.108	3	2.036	12.37	< .01
Mean	.745	3	.248	17.95	< .01

Halo Analysis[b]

Source	Sums of Squares	DF	Mean Square	F	p
Scale Format	7.905	3	2.635	164.38	< .01

Range Restriction Analyses[b]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	8.01	3	2.67	59.66	< .01
2	3.67	3	1.22	26.66	< .01
3	4.14	3	1.38	26.37	< .01
4	5.97	3	1.99	50.27	< .01
5	4.77	3	1.59	24.49	< .01
6	3.61	3	1.20	24.52	< .01
7	4.86	3	1.62	35.19	< .01
8	5.34	3	1.78	36.28	< .01
9	4.18	3	1.39	23.71	< .01
Mean	4.53	3	1.51	152.22	< .01

[a] AS < MSS < GRS = BARS

[b] AS > MSS = GRS = BARS

Source Tables for Main Effect of Ratee:
Modified Grid Form of REP Test

Leniency Analyses[a]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	130.92	2	65.46	218.39	< .01
2	143.36	2	71.68	264.68	< .01
3	94.45	2	47.23	126.69	< .01
4	102.90	2	51.45	143.77	< .01
5	181.59	2	90.80	280.34	< .01
6	114.94	2	57.47	172.11	< .01
7	114.40	2	57.19	172.11	< .01
8	122.92	2	61.46	162.76	< .01
9	90.71	2	45.35	96.71	< .01

Halo Analysis[a]

Source	Sums of Squares	DF	Mean Square	F	p
Ratees	2.91	2	1.46	43.30	< .01

Range Restriction Analyses

Ratees were not a separate effect for the Range Restriction analyses since these were based on the standard deviations across ratees as data points.

[a]In all cases, the standard deviations were greatest for the "worst" instructor, intermediate for the General Psychology instructor, and lowest for the "best" instructor.

Source Tables for Main Effect of Ratees:
Factor Analysis of REP Test

Leniency Analyses[a]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	130.92	2	65.46	206.82	<.01
2	143.36	2	71.68	248.89	< .01
3	94.45	2	47.23	127.78	<.01
4	102.90	2	51.45	155.49	<.01
5	181.59	2	90.80	267.68	<.01
6	114.94	2	57.47	169.53	<.01
7	114.40	2	57.19	173.06	<.01
8	122.92	2	61.46	164.20	<.01
9	90.71	2	45.35	109.01	<.01

Halo Analysis[a]

Source	Sums of Squares	DF	Mean Square	F	p
Ratees	2.91	2	1.456	42.42	<.01

Range Restriction Analyses

Ratees were not a separate effect for Range Restriction analyses since these were based on the standard deviaitons across ratees as data points.

[a]In all cases, the standard deviations were greatest for the "worst" instructor, intermediate for the General Psychology instructor, and lowest for the "best" instructor.

Source Tables for Main Effect of Ratees:
Sorting Task

Leniency Analyses[a]

Dimension	Sums of Squares	DF	Mean Square	F	p
1	130.92	2	65.46	215.21	<.01
2	143.36	2	71. 8	261.52	<.01
3	9 .45	2	47.23	124.92	<.01
4	102.90	2	51.45	142.53	<.01
5	181.59	2	90.80	271.97	<.01
6	114.94	2	57.47	168.58	<.01
7	114.40	2	57.19	171.73	<.01
8	122.92	2	61.46	164.04	<.01
9	90.71	2	45.35	96.93	<.01

Halo Analysis[a]

Source	Sums of Squares	DF	Mean Square	F	p
Ratees	2.91	2	1.456	42.81	<.01

Range Restriction Analyses

Ratees were not a separate effect for Range Restriction analyses since these were based on the standard deviations across ratees as data points.

[a]In all cases, the standard deviations were greatest for the "worst" instructor, intermediate for the General Psychology instructor, and lowest for the "best" instructor.

THE RELATIONSHIP BETWEEN COGNITIVE COMPLEXITY AND THE USE OF VARIOUS TYPES OF RATING SCALE FORMATS

by

MARY ANNE LAHEY

B. S., Illinois State University, 1976

AN ABSTRACT OF A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Psychology

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1980

Compatibility between raters' cognitive complexity, the degree to which persons are able to discriminate among dimensions of complex stimuli, and the cognitive demands of rating scale formats has been shown to aid in the reduction of undesirable psychometric properties of performance appraisal ratings (Schneier, 1977). Although these findings appear to be encouraging in the determination of rater characteristics that may affect ratings, failures to replicate this study have raised troubling theoretical and practical issues.

This study further investigated the cognitive complexity-rating scale format relationship by attempting to identify the skills or abilities that comprise a complex, as opposed to a simple, rater, and also by attempting to identify the particular aspects of rating scales that render them complex or simple.

The cognitive complexity of 96 undergraduate students was measured using each of three different measuring instruments chosen to represent important aspects of the complexity construct: a) a modified grid form of the Role Constructs Repertory (REP) test; b) a factor-analytic scoring technique for the original REP test; and, c) a sorting task measure of complexity. Performance ratings for three of the students' instructors (their General Psychology instructor, their "best" instructor for that semester, and their "worst" instructor for that semester) were obtained using each of four rating scale formats presumed to vary in complexity: a) a 7-point behaviorally anchored rating scale; b) a mixed standard rating scale; c) a 7-point graphic rating scale; and, d) a 3-point "simple" rating scale. Each of these scales tapped the same nine performance dimensions.

High- and low-complexity rater groups were formed by splitting complexity scores at the median. Ratings from these groups were compared with respect to leniency, halo, and restriction of range. Separate analyses were performed for each of the three complexity measures.

Leniency analyses were based on Complexity x Ratee x Scale Format ANOVAs for each of the nine rating dimensions. No main effects of Complexity, nor any Complexity x Scale Format interactions, emerged for any of the dimensions or for any of the complexity measures.

Complexity x Ratee x Scale Format ANOVAs, using standard deviations across dimensions as data points, also supplied halo information. No significant Complexity main effects nor Complexity x Scale Format interactions were found for any of the complexity measures.

Range restriction analyses were based on Complexity x Scale Format ANOVAs, using the standard deviations across ratees as data points for each dimension. Although a slight trend toward differences between the rater groups emerged (cognitively complex raters tended to provide ratings with less range restriction), these results only reached significance for a small number of dimensions. No significant Complexity x Scale Format interactions were found for any of the complexity measures.

The results of this study, therefore, failed to support the hypothesis that compatibility between raters' cognitive complexity and the cognitive demands of rating scale formats is a moderator of behavior. Supported by findings of weak relationships among the complexity measures employed here, the obscure nature of the concept of cognitive complexity is proposed as one reason for lack of consistency among research results in this area.

Other potential sources of confusion in the determination of the cognitive complexity-rating scale relationship lie in the rating scales themselves. Use of those scale formats commonly found in the current performance appraisal literature failed to shed light on the characteristics of rating scales that render them complex or simple. It is argued, however, that the number of dimensions tapped by a rating scale may be an important variable.

Overall, it appears that any moderating role that a rater's cognitive complexity may play in affecting rating behavior is more complex than a compatibility hypothesis would predict. If a relationship does exist between cognitive- and rating scale-complexity (and support for this contention is dwindling), further investigations are necessary to determine its nature.