INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of “sectioning” the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
Finnegan Schaalman, Mary Lou

BEHAVIORAL CLARIFICATION OF RATING DIMENSIONS

The Ohio State University

University Microfilms International 300 N. Zeeb Road, Ann Arbor, MI 48106

Ph.D. 1984
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy.
Problems encountered with this document have been identified here with a check mark. √

1. Glossy photographs or pages ______
2. Colored illustrations, paper or print ______
3. Photographs with dark background ______
4. Illustrations are poor copy ______
5. Pages with black marks, not original copy ______
6. Print shows through as there is text on both sides of page ______
7. Indistinct, broken or small print on several pages √
8. Print exceeds margin requirements ______
9. Tightly bound copy with print lost in spine ______
10. Computer printout pages with indistinct print ______
11. Page(s) _______ lacking when material received, and not available from school or author.
12. Page(s) _______ seem to be missing in numbering only as text follows.
13. Two pages numbered _______. Text follows.
14. Curling and wrinkled pages ______
15. Other ____________________________________________________________
BEHAVIORAL CLARIFICATION OF RATING DIMENSIONS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree of doctor of Philosophy in the Graduate School of The Ohio State University

By

Mary Lou Finnegan Schaalman, B.A., M.A.

****************

The Ohio State University

1984

Reading Committee
Milton D. Hakel
Robert S. Billings
Richard J. Klimoski

Approved by
Milton D. Hakel
Advisor
Department of Psychology
ACKNOWLEDGMENTS

The U.S. Navy cooperation with this project was greatly appreciated. That cooperation was approved by Thomas D. Williams, Commander, USN, at the Leadership and Management Education Branch in Washington, D.C. Given his approval, this project received support from the Officer Indoctrination School staff at Newport, Rhode Island, under the direction of Commander Balling and Lieutenant Commander Hess. Lieutenant Sara Leeds and Lieutenant Ron Mase provided assistance in scheduling and coordinating the new officers' time in order to involve them in this project.

Several of my colleagues at McBer and Company contributed invaluable assistance by acting as liaisons with the Navy and by providing technical and, especially, moral support. Thanks go to all those who encouraged this effort and made it possible for me to adjust my work schedule to meet the challenge of this project. Particular thanks go to Wayne Smegov, Mike Sokol, Susan Ennis, Dawn Jabari, Elisse Winer, and especially to Bernard Cullen and Jane Lawson.

I am also grateful for the insight, rational perspectives, and friendship of Joan Brett and Paul Sackett. Joan in particular provided a base of support and a continuing link with the Industrial/Organizational Psychology program at Ohio State.

Despite the protracted struggle over several years, the committee members were remarkably patient and, in the long run, supportive of my continuing efforts to complete this project. Very special thanks go to Milt Hakel. He advised, calmed and encouraged me at times when I needed each of those. His intellectual gifts and pragmatic perspective were invaluable resources for me. His example and his friendship helped get me through. An added benefit of working with him was getting to know Lee Hakel better. I owe her thanks for her hospitality and for her tolerance of the many requests that I made of Milt that were inconvenient for them both.

Last but not least, I am exceedingly grateful for the patient, undemanding, yet unhesitating support of my mother Anne W. Finnegan and my husband Don Schaalman.
VITA

May 2, 1953.......................... Born - Boston, Massachusetts

1975................................. B.A. Jackson College of Tufts University; Medford, Massachusetts

1977................................. M.A. The Ohio State University; Columbus, Ohio

1978................................. Lecturer, Department of Psychology, Tufts University; Medford, Massachusetts

1978-1984........................... Senior Associate, McBer and Company; Boston, Massachusetts

PUBLICATIONS


FIELDS OF STUDY

FRENCH (B.A. cum laude)
Emphasis on 16th and 20th century theatre.

PSYCHOLOGY (B.A. magna cum laude)
Emphasis on social psychology including conflict and role theory.
Advisor: Professor Jeffrey Z. Rubin, Ph.D.

INDUSTRIAL/ORGANIZATIONAL PSYCHOLOGY (M.A.)
Emphasis on motivation and on measurement and legal issues in performance assessment.
Thesis focused on decision making and probability estimation.
Advisor: Robert S. Billings, Ph.D.

INDUSTRIAL RELATIONS (Doctoral-level minor)
Emphasis on collective bargaining and the regulatory environment.

INDUSTRIAL/ORGANIZATIONAL PSYCHOLOGY (Ph.D.)
Dissertation focused on measurement issues and cognitive aspects of performance assessment.
Advisor: Professor Milton D. Hakei, Ph.D.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENT</td>
<td>ii</td>
</tr>
<tr>
<td>VITA</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>PROLOGUE</td>
<td>ix</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. CONCEPTUAL &amp; THEORETICAL BACKGROUND</td>
<td>4</td>
</tr>
<tr>
<td>III. TRANSLATING CONCEPTS INTO OPERATIONS</td>
<td>41</td>
</tr>
<tr>
<td>IV. METHOD</td>
<td>54</td>
</tr>
<tr>
<td>IV. ANALYSES</td>
<td>64</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>68</td>
</tr>
<tr>
<td>V. DISCUSSION</td>
<td>97</td>
</tr>
<tr>
<td>EPILOGUE</td>
<td>124</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>125</td>
</tr>
<tr>
<td><strong>APPENDICES</strong></td>
<td></td>
</tr>
<tr>
<td>A. DEFINITIONS</td>
<td>130</td>
</tr>
<tr>
<td>B. FILM SYNOPSIS</td>
<td>131</td>
</tr>
<tr>
<td>C. RATING DIMENSION CHOICE</td>
<td>133</td>
</tr>
<tr>
<td>D. PILOT RATING DATA</td>
<td>134</td>
</tr>
<tr>
<td>E. MATERIALS</td>
<td>135</td>
</tr>
<tr>
<td>F. MEAN DIMENSION SCORES BY RATEE</td>
<td>158</td>
</tr>
<tr>
<td>G. INTERCORRELATION OF DIMENSION RATINGS</td>
<td>159</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>H. CONTENT CODING CATEGORIES AND RULES</td>
<td>163</td>
</tr>
<tr>
<td>I. EXAMPLES OF CODED PASSAGES</td>
<td>164</td>
</tr>
<tr>
<td>J. COMPARISON OF SCORES FROM CONDITION 3 &amp; 4</td>
<td>165</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Distribution of the dependent variable for ANOVA: Standard deviation of ratings</td>
<td>69</td>
</tr>
<tr>
<td>2. ANOVA for HALO defined as s.d. of ratings: Conditions 1, 2, 3</td>
<td>70</td>
</tr>
<tr>
<td>3. A comparison of means for the DV: s.d. of ratings</td>
<td>70</td>
</tr>
<tr>
<td>4. Hypothesis 1: The effects of behavioral clarification of dimensions on the subsequent HALO among dimension ratings</td>
<td>72</td>
</tr>
<tr>
<td>5. Hypothesis 2: The effects of behavioral clarification of dimensions on subsequent interrater reliability of dimension ratings</td>
<td>74</td>
</tr>
<tr>
<td>6. Demographic profile of Ss by condition</td>
<td>77</td>
</tr>
<tr>
<td>7. Responses to the manipulation check questions</td>
<td>80</td>
</tr>
<tr>
<td>8. Responses to the manipulation check questions: Condition comparisons by t-test</td>
<td>80</td>
</tr>
<tr>
<td>9. ANOVA for Rater experience</td>
<td>81</td>
</tr>
<tr>
<td>10. ANOVA for Perception that list helped</td>
<td>81</td>
</tr>
<tr>
<td>11. ANOVA for Feeling prepared</td>
<td>82</td>
</tr>
<tr>
<td>12. Anova for Knowing to which dimension the behavior was related</td>
<td>82</td>
</tr>
<tr>
<td>13. ANOVA for Dimensions refer to different characteristics</td>
<td>83</td>
</tr>
</tbody>
</table>
14. Forced rankings of rater choices to improve their rating ....................... 86

15. Content coding: The nature of the data raters report considering when making ratings .......... 90

16. ANOVA for Content coding of dimension definitions:
   # of words ........................................... 90

17. ANOVA for Content coding of dimension definitions:
   # of action statements ............................. 91

18. ANOVA for Content coding of dimension definitions:
   Specific actual behavior referenced .............. 91

19. ANOVA for Content coding of dimension definitions:
   Refers to outcomes .................................. 92

20. ANOVA for Content coding of dimension definitions:
    Non-behavioral statement of construct ............ 92

21. HS2: Scores from Ss in Condition 3 will not differ from scores from Ss in Condition 4--
    HALO.................................................. 95

22. HS2: Scores from Ss in Condition 3 will not differ from scores from Ss in Condition 4--
    INTRERRATER RELIABILITY ......................... 95
A company needed to fill a new supervisory job in the production department of its Plant X. Management decided to carry out a group selection using assessment center exercises and a panel of observers to assess important dimensions of the candidates' performance. One of the key qualities for performance of the new job was initiative. The production department was reorganizing and the person in the target job would be dealing with and making changes to help the new configuration work. Four of the candidates received the following ratings on initiative.

- Stan came from Plant Y which was organized in a fashion similar to the new plan for the Plant X production department. His actions and ideas reflected a commitment to, understanding of, and significant level of comfort with operating in the Plant Y style. Stan was rated high on initiative.

- Jean charged into conversations and made suggestions, but she was not receptive to others' ideas. She stood her ground and argued. One of the assessors talked the others into lowering the consensus rating for Jean on initiative because he thought she had too much of that quality. Jean was rated moderately high on initiative.

- Dennis had been assigned to the corporate office on a special project for a year. Under the direction of a corporate vice president he was working on an innovative new inventory control program. In the assessment exercises Dennis's lack of production experience showed: his presentations and ideas were creative but impractical. Dennis was rated high on initiative.

- Tom made a point of being the first to speak in each exercise—he volunteered to go first. He suggested new ideas or ways to approach problems. The assessors thought he "must have been told that to go first was a good idea," and they got tired of always hearing from him first. They didn't believe he showed initiative on his own. Tom was rated very low on initiative.
INTRODUCTION

In the assessment of observed behavior, an individual is typically asked to rate the degree to which another person has demonstrated a characteristic (e.g., initiative) considered to be an important dimension of the person's job performance; the observer makes the ratings after first witnessing how the person to be rated interacts with others or performs tasks in the course of doing his/her job. As noted by Finkle and Jones (1970) "among the most basic requirements for the profitable generation and utilization of assessment information is the development of a language in which judgments can be formulated, weighed against other evidence, and communicated with reasonable accuracy and understanding" (p. 67). They go on to argue that the characteristics being assessed (e.g., initiative) must be carefully defined so that all the people making assessments interpret and judge in the same way when evaluating individuals in terms of characteristics important for performance.

Finkle and Jones (1970) make a practical suggestion to improve the definition and reduce the ambiguity of assessment dimensions. Their concern is mirrored in measurement theory (c.f. Ghiselli, Wherry, and others)
and in the academic research literature that seeks to explain and improve a range of industrial and organizational applications of the measurement of psychological variables. This research explores some of the principles of measurement theory with an eye to improving the application of that theory in the assessment of human performance on (or for) the job.

The following research has evolved from the basic proposition that clarification of assessment dimensions in concrete behavioral terms will enhance (1) the degree to which raters differentiate among assessment dimensions when rating (i.e., reduction of halo) and (2) the interrater reliability of those ratings. Whereas most research has attempted to affect these psychometric properties of ratings by clarifying points on the scales used to rate dimensions, this effort focuses on a different issue: the clarification of the rating dimensions as constructs representing aspects of performance that can be observed in job behavior. The derivation of this research effort from measurement theory and its relevance to applied research will be explained in the following pages.

This document begins with a presentation of the conceptual and theoretical background of the research
issue. After beginning on a general plane, the discussion shifts to consideration of the research focus on behavioral clarification of assessment dimensions. Following this conceptual and theoretical background, the document further explains the rationale for the choices made in translating concepts into the operational characteristics of the study. The usual Method, Results, and Discussion sections follow in order.
CONCEPTUAL AND THEORETICAL BACKGROUND

In everyday life, virtually all people assess others' behavior—whether to figure out why a person did something or to anticipate his/her next move. This happens in a chess game, in a family, in a classroom, a clinic or an office. Whenever one person attempts to determine the actions, thoughts, or motives of another person, the first person is actually making psychological assessments of the other in terms of characteristics on which people can vary.

In organizations, interpersonal assessments are often formalized into ratings. As part of their job duties, supervisors typically rate individuals on a variety of dimensions chosen to reflect characteristics which the organization defines as important for performing the job. The rating process involves a complex response by the rater. "An accurate rating is a function of three major components: performance by the ratee, observation of that performance by the rater, and the recall of those observations by the rater" (Wherry & Bartlett, 1982; p.521).

Wherry's theory of rating conceptualizes the important aspects of ratings and the rating process. He applied that theory in his theoretical writings and further espoused that theory in theorems and prescriptions for
practice. Throughout his career, Wherry was noted for his insistence that "you can draw much more objective conclusions if you stick to descriptive, observable evidence" (Bartlett, 1982 p. 517). He saw this as a practical maxim in keeping with the theoretical and empirical components of the theory of rating.

Unfortunately, many of the psychological variables of interest to researchers, managers, teachers and clinicians are not tangible. In other words they are not immediately observable or measurable. But though such variables are not concrete, they are nonetheless often deemed to be important characteristics for performance, and therefore need to be measured. Thus a means of measuring these intangible characteristics must be devised—and the means is perforce indirect. Ghiselli, Campbell, and Zedeck (1981) note that "it is doubtful that we could find any psychological variable of interest for which there would be total acceptance with respect to its measurement or definition" (p. 10). Disagreement—implicit and explicit—about variable definitions and measures has consequences for the quality and comparability of ratings of those variables.

Take as a simplified example two variables: height and intelligence. Raters are asked to observe a person on
the job and then to rate the person on these two variables. Height has both a standard meaning, commonly held and a standard measurement approach, commonly employed. Although the raters' ratings might not be entirely accurate, the raters would be expected to agree reasonably well because of the shared definition that implicitly directed them to observe the same tangible characteristic of the person. Discrepancies would arise only from the raters' differences in recall, not from differences of what was observed, nor from any changes in the ratee's height. In fact, in a study conducted by Dawes (1977) height measured using a standard ruled measure was correlated with height as measured using five different rating scale formats common to psychological research; correlations ranged from .88 to .94. By contrast ratings of intelligence would probably be more variable. The ratee's behavior that is thought to be related to intelligence (e.g., manner of speech, ideas, actions) would inevitably change during the observation, causing the raters to observe different things and recall different things. In short the ratings would probably differ greatly unless the variable intelligence was initially clarified in terms that made it less intangible and more concrete, observable, and measurable.
Some psychological assessment is done by means of structured paper-and-pencil instruments which clearly maintain a constant definition and measurement scheme for the variable(s) in question. Even though those instruments may measure self reports or other indirect evidence of a characteristic, the structured format and scoring procedures operationalize a single definition, which enables measurements to be compared: assessors can use the standardized measure to assess one definition of a variable in the same way for all ratees.

A great deal of assessment, however, occurs in classrooms, clinics, and offices, where it typically takes the form of (1) the observation of behavior and (2) the rating of the degree to which some intangible construct was demonstrated in that behavior. Individuals are constantly making assessments by drawing inferences from the behavior they observe. Whereas in a paper-and-pencil inventory the choice of items essentially defines what is to be considered relevant to the assessment dimension to be measured, in the practice of rating observed behavior, definitions of performance-related variables are not always provided for the rater. It is often unclear which observations are relevant to the measurements. Sometimes raters are simply given names for rating dimensions. In
other circumstances raters recieve conceptual definitions which each rater must translate for him/herself into lists of behavioral observations relevant to the different dimensions.

The name given to an assessment dimension--a trait, or other variable of human behavior--is usually carefully selected to give a reasonable representation of the definition of that variable. Nonetheless, trait names and other variable labels are used with different connotations by each person who applies them. Take, for example, the variable need for achievement (nAch). McClelland originally defined nAch as a "social motive", that is a predisposition revealed in a person's thinking, that results in his/her selection of behavior which leads to unique achievement, efficiency, and high standards. His choice of the TAT as a measure was consistent with the definition of nAch as a motive. He and several other researchers who accepted his definition and method of measurement have published numerous studies demonstrating the construct validity of nAch as measured by the TAT and its utility for predicting certain conceptually related behaviors. Other researchers have developed measures (e.g., Edwards's Personal Preference Schedule) that operationalize nAch as "values achievement"; these respondent-format instruments assess whether a person
values achievement, efficiency and high standards. The respondent-inventory measures of nAch do not show as strong a relationship with actual achievements as does the TAT measure. Thus, although the two measures share the name of their variable and part of a definition, they are defined differently and they measure differently, tapping different aspects of the definition, and thereby resulting in different levels of validity for predicting behavior (McClelland, 1980). A person may value achievement but not think in ways or select behaviors that help him or her to achieve. In sum, as Ghiselli et al. (1981) note, we must be careful about comparing results or measures of traits that are nominally the same.

Clearly in the case of nAch, Ghiselli et al.'s (1981) point is well taken; at both the conceptual and the operational stages of development, there is not likely to be an easy acceptance of one definition or measurement strategy for intangible psychological variables. Respondent measures of such variables may evolve through discussion and controversy, but once the format, items, and scoring rules are in place, all measures are made with reference to the variable as defined in the measure. Since much assessment is done after observation of operant behavior, similar clarity of definition would be useful in order to assure conceptual and empirical comparability of
ratings. Operant behavior provides a wealth of information on individual performance that is inaccessible to respondent measures (i.e., self-report responses to a limited set of questions focused on a subset of all possible performance behaviors). However, idiosyncratic interpretation and rating by assessors of observed behavior diminish the potential advantages of observer ratings over self reports. The point here is not to explore or contrast the two types of ratings, but rather to suggest that one of the types (observation) could profit by adopting some of the rigor of the other (respondent measures). The definitions of rating dimensions should be clearly articulated so that all users are referring to the same quality when they rate "initiative" or "sociability" or "nAch" or any other dimension of performance. Although the preceding statement seems like common sense and is widely prescribed (see: Standards and Ethical Considerations for Assessment Center Operations, 1980) it is often neglected in operation. Assessment dimension definitions and examples provided in initial training efforts eventually become folklore, and the meanings are all assumed instead of spelled out (Williamson & Schaalman, 1980). Whereas a strong norm may firmly establish a definition and examples, new raters may not share these definitions. The definitions need, therefore, to be presented to new raters
along with clear examples in order to prevent noncomparable interpretations by different raters.

The conceptual and operational ambiguity of assessment dimensions that are used to rate performance is assumed to result in idiosyncratic interpretations among raters of the observed behaviors. The anticipated consequences of reducing that ambiguity—that is of standardizing the operational definitions of rating dimensions—would be two-fold.

(1) Raters would be better able to sort observations, behaviors, and impressions into a set of conceptually distinct performance dimensions. Differentiation among various aspects of observed performance and among the subsequent dimension ratings for a ratee, would result in an increase in the empirical independence of ratee's ratings (i.e., a reduction in halo).

(2) By sharing a definition of the rating dimension, agreement among raters should increase. Empirically, interrater reliability should increase as different assessors rate the same ratees and behaviors, using a shared conceptual and operational definition of a dimension.

The Rationale for Behavioral Clarification
of Assessment Dimensions

How should the clarification of performance dimensions be accomplished? Various possibilities exist. The focus
of this discussion is on observer ratings of intangible constructs and the research design tests the effects of clarifying dimensions in terms of behavioral examples. The treatments (1) clarify the constructs which underlie the rating dimensions by providing behavioral referents for the identification of observations relevant to each dimension, and (2) manipulate the raters' use of those behavioral referents that clarify the constructs. The conceptual and empirical rationale for this design is outlined in the following discussion.

In his seminal 1938 work on the multimethod approach to the assessment of personality, Murray argued that subjective assessment is a flawed means of defining the complex configuration of variables that can be used to describe an individual's personality. He also suggested that assessors need to observe a multitude of behaviors, over time, to get the best picture of personality. Murray's ideas were instrumental, a few years later, in the design, operation and evaluation of the now famous OSS assessment center. He envisioned the systematic observation of behavior as a means to clarify subjective assessment of the complex variables that make up successful job performance. Since that time, everincreasing numbers of persons have rated, and been rated on, performance dimensions following the observation
of behavior. In those years considerable research has been conducted to explore and explain the merits and problems of dimension ratings of observed performance.

In 1952 Wherry first published his theory of rating. Wherry systematically proposed a set of formulae detailing the components of ratings based on observation of a ratee. He also drew conclusions and postulated theorems in keeping with both the empirical premises of the theory and knowledge of the nature of human observation, recall, and judgment. The focus of many of his theorems is the rater. Since the rater is key in two of the three components of the basic rating equation, it is not surprising that prescriptions for rating accuracy suggest ways to help the rater to focus attention on important aspects of performance. This research tests one approach to helping raters accurately rate important aspects of performance.

Wherry suggested that although rater recall is partially based on observation of the rater's actual performance, "memory is more than the sum of its parts. It is partly, sometimes largely, composed of an imposed schema or systematic bias factor of recall" (Wherry &

*Rating = (ratee performance + error) + (rater observation + error) + (rater recall + error)
Plausible but invalid assumptions may lead raters to see behaviors as related to a dimension that was never intended to subsume them; implicit assumptions may lead raters to see correlations between rating dimensions that have no correlation. Such misjudgments result from reliance on personal interpretations instead of on standard definitions and examples. In reviewing the literature on halo, Cooper (1981) discussed a number of studies in which implicit personality theories about the relationships among different aspects of behavior were shown to explain suspiciously high intercorrelations of dimension ratings.

Likewise the research literature on the factor structure of assessment center ratings has suggested that overall assessment ratings can be accurately modeled with a small subset of dimension ratings (Sackett & Hakel, 1979). The factor analysis revealed what seemed to be a simple schema in use by assessors. Sackett and Hakel concluded that instruction to use all dimensions is inappropriate, since their analysis showed that the dimensions do not all represent identifiable, useful information. But is it true that the dimensions do not represent useful information, or are they simply presented and used in ways that are
ineffective for reliably identifying and distinguishing true dimensions of performance? Perhaps the factor structure represents the true minimum dimensionality of performance or perhaps it represents the instructions given to assessors. The norm is to tell assessors that an exercise will primarily provide information on a limited number of identified dimensions but that all dimensions may be shown. The question then arises: if assessors were given clear, distinct definitions of dimensions would they attend to a broader range of behaviors and agree more on the importance of the various dimensions?

In their exposition of rating theory Wherry and Bartlett (1982) state the following theorem.

**Theorem 5:** Rating scale items which refer to easily observed behavior categories will result in more accurate ratings than will those which refer to hard-to-observe behaviors.

Similarly, Dunnette (1966) contends that the most pervasive source of error in behavior ratings is the design of rating instruments in which behavior descriptions are not understandable, relevant, or practical. Consequently, unreliable or undifferentiated ratings may emerge, because the characteristic being assessed is abstract, stated in ways that are remote from real-life behavior, and accordingly very difficult to consider in a consistent manner (Finkle & Jones, 1970; p.
One way to make assessment dimensions less abstract, more easily understood, and more easily applied would be to provide raters with clear specific statements of what behaviors are relevant to each of the constructs to be rated.

Another common tendency among raters is to give highly inconsistent ratings, apparently without regard to the behavioral content of items (Dunnette, 1966; p 90). Because of this tendency, researchers have often attempted to improve performance ratings in the way suggested by Wherry (1982). They have interpreted Wherry's theorem as suggesting that by making the points on the rating scales that are used to measure dimensions more explicit and behavioral, rating accuracy can be improved.

In a study conducted by Barrett, Taylor, Parker and Martens (1958), ratings from unanchored scales labeled simply with trait names were compared with trait ratings from scales anchored with behavioral examples at various scale points. The latter ratings proved to be more reliable and had less halo. In subsequent years, numerous studies and scale-construction methods have been analyzed in an attempt to verify that behavioral clarification of scale points improves rating reliability and reduces halo.

The evidence from these studies indicates that, yes, behavioral clarification of dimensions can improve the
psychometric properties of ratings (e.g., reduce halo). However, clarification need not take the form of behavioral anchors at scale points. One of the most widely researched approaches to behavioral specification of scale points is the BARS method (Smith & Kendall 1963). After reviewing the research on BARS, Landy and Farr (1980) concluded that BARS scales have not been shown to outperform other scales in terms of halo reduction. The works of Friedman and Cornelius (1976) and of Warmke and Billings (1979) suggest that the superior psychometric qualities of scales may derive from participation and training of raters. Warmke and Billings (1979) report reduction of halo errors for a treatment group which participated in rating-scale construction and no such effect for a control group or for groups either receiving a lecture or participating in a discussion focused on the nature of rating errors to be avoided. Although Warmke and Billings chose to interpret their findings in terms of increase in understanding of scale use and in motivation for scale use, they briefly refer to the possible explanation of improved understanding of the job components rated. This explanation is reasonable in the context of the preceding discussion of clarification of dimensions as a possible means to improve differentiation
of dimensions by assessors. The scale construction
treatment is the only place where the authors describe
"clarifying and behaviorally defining dimensions" (p 126).
In consideration of the preceding body of research, it
seems logical to test the idea that raters who are
provided with information on which behaviors are relevant
to which assessment dimensions are more likely to
differentiate among dimensions when rating, and to agree
with one another's ratings.

Wherry and Bartlett (1982) suggest a theorem in the
theory of rating that "the rater will make more accurate
ratings when he has been forewarned concerning the types
of activity to be rated since this will facilitate his
more properly focusing attention on such pertinent
behavior" (Theorem 6; p. 533). Research has demonstrated
the merit of this proposition. For example, on the basis
of his 1979 review of biases in the interviewing process
Arvey recommends clear articulation of job requirements as
a prerequisite for improving reliability and validity of
assessments made by an interviewer. In a later review
written with Campion (1982), Arvey reports a study in
which impressive increases in reliability and validity of
assessments were associated with building a structured
interview around behaviors shown to be related to job
success. Despite the questionable interview format (i.e.,
hypothetical questions would have tapped the interviewee's "espoused theory," not necessarily his/her "theory in use,"), the clarification of relevant behaviors for performance assessment seems to have contributed to the increased interrater reliability attained.

It has been suggested that the use of the BARS (or BES) rating format may produce ratings which are psychometrically superior to ratings made on scales of other formats (Smith and Kendall, 1963). In their 1975 review of the BARS research literature, Schwab, Heneman, and DeCotiis concluded that most research on BARS up until that point had not compared them to other types of scales; those studies that had had produced mixed results with respect to leniency, dimension independence, and reliability of ratings. In a later comparison of BARS and summated ratings, Bernardin, Alvares and Cranny (1976) report mixed results in terms of the interrater reliability of the two types of ratings. The summated scales were constructed by identifying behaviors related to the dimension to be assessed; the behaviors were each rated on a Likert scale and the ratings summed for a final scale score. Both the BARS format and the summated scales use a detailed description of behavior and the behavior is strictly assigned to a scale. Perhaps it is this strict
attention paid to rating in terms of dimension related behaviors that is instrumental in determining the levels of halo and interrater agreement in ratings. By focusing on the sophistication of scaling techniques, researchers may have overlooked the fundamental reason for the otherwise seemingly variable differences between BARS ratings and ratings from other formats. In a 1977 study (Bernardin) comparing BES (i.e. BARS) to summated scales, items generated by the BES process for item development were also used as items for summated scales. No differences in interrater reliability or halo were found across formats.

The BARS approach to rating scale development formally combines some of the rules which make up the common wisdom for developing the best rating scales. It involves users in the development of an appraisal instrument that reflects important aspects of a job, describing those aspects in the language of job incumbents. It also applies a fairly rigorous scaling method for placing behaviors on the rating scales. Although it has not addressed the behavioral focus in the way this study has, the research on BARS has often suggested methods for improving BARS ratings by means which essentially require a greater attention to the behavior of the person being rated.
After reviewing the many variations on the basic BARS development process Bernardin et al. (1976) came up with a set of prescriptions for the best approach. Among their recommendations was the suggestion that raters place critical incident examples on the scales along with the ratings they make. In essence this suggests that raters justify each rating by citing the actual observed behavior on which it was based. A year later in 1977, Bernardin and Walter published another study of BARS in which they compared a variety of training programs and scale formats. The treatment group which produced ratings with the least halo and best interrater reliability was one which not only was trained, BUT ALSO kept diaries in which they recorded scale relevant behavior as it occurred. The first of the two preceding articles suggests, then the latter confirms the merits of requiring raters to observe and record discrete behavior relevant to ratings that will be made. Both raters using BARS in this recommended way and those using summated scales make an increased number of judgements per scale. The predicted outcome is less halo and greater reliability of ratings from raters who use a checklist (similar in concept to keeping a diary list of behaviors for each scale) than for raters who are less sure about what behaviors relate to a scale or who have not carefully recorded the behaviors observed.
In a review of the literature on performance rating and training of raters (1981), Latham and Wexley draw conclusions consistent with this proposition that behavioral clarification of the dimensions used to rate observed performance is likely to improve the psychometric properties of the ratings. In reviewing a program of research conducted by Bernardin and his colleagues they conclude that "rater training programs, if they are to be effective, should concentrate on enhancing the accuracy of ratings through the discussion of the multidimensionality of work performance, the importance of recording objectively what is seen, and the development of specific examples . . . ." (p 106). Finally they conclude that only training programs that give trainees a chance to practice observing and rating behaviors—*with feedback on the accurate application of dimension ratings to those behaviors*—are likely to be effective in improving rating accuracy. Although Latham and Wexley's emphasis is on the feedback of rater accuracy, it is also possible in the context of the discussion to this point, to suggest that behavioral definition of what is relevant to dimension ratings is an important contributor to the success of those training efforts.

In his theory of rating Wherry also contends that any factor that will heighten sensory experience with the
ratee and his/her performance will increase the influence of actually observed behavior upon ratings made, as against the influence of bias or error in the act of percieving. Thus, he postulates the following theorem which is particularly relevant to the research at hand.

Theorem 8: If the perceiver is furnished an easily accessible checklist of objective cues for the evaluation of performance to which he can frequently refer, he should be better able to focus his attention properly. (1982; p. 534)

One study that clearly demonstrates the merit of this proposition was conducted by Newcomb in 1931. He conducted studies of boys at summer camp, in which counselors completed daily checklists of observations on 26 behaviors. The counselors also rated the boys at the end of the camp session. Newcomb found that the intercategory correlations among the ratings made at the end of the session were substantially higher than the correlations derived from daily checklist observations (Cooper, 1981). Newcomb interpreted these findings as indicative of a halo effect in the ratings. He presumed that the close relationship between ratings sprang from logical presumptions in the minds of the raters rather than from actual behaviors (Newcomb, p. 228). Newcomb's findings are similar to Bernardin and Walter's (1977)
findings concerning the use of diaries to keep track of observed behavior and the subsequent psychometric properties of those ratings. The use of behavior lists to organize one's perceptions of observed behavior, and to record those observations, appears to facilitate differentiation of dimensions, and one could predict, therefore, that use of such lists increases interrater reliability. Dreher and Sackett (1983) note the potential advantages of the checklist approach to making assessment center ratings:

Assessment procedures that attempt to score exercises objectively or use behavioral checklists or other means of making an explicit connection between a specific behavior and the rating given for a particular dimension would seem to have a much greater likelihood of being consistent with the Guidelines than procedures that rely on global judgments by assessors following observation. (p. 254)

The use of checklists as instruments for measuring job performance began before Wherry first published his rating theory (1938). Thurstone (1927, 1928) and Richardson and Kuder (1933) began by generating lists of task statements and experimenting with scoring methods. Studies compared (a) ratings from groups of raters who checked performance of each item with (b) ratings from raters who indicated the relative frequency with which each item was performed (e.g. always/usually/sometimes/seldom/never) (Ferguson, 1947). Checklists appeared to have considerable potential for use in assessment of
performance since they were fast and easy to score, required only simple instructions to raters, and were quite reliable \((r=0.80, \text{Knauf}, 1948; r=0.95-0.97, \text{Uhrbrock}, 1950)\). Those who studied checklists in the 1950s continued to focus on empirical improvements in scoring techniques \((\text{Jurgensen 1949; Uhrbrock 1950, 1961; Meyer 1961; Siegel 1954; Prien and Campbell 1957})\).

Nonetheless, checklists seemed to fall out of fashion in the 1960s.

While it is not entirely obvious why checklists did not grow in favor for performance appraisal purposes, a few expalanations seem reasonable. First of all, they continue to be used often as part of job analyses which serve as the bases of compensation systems. Checklists, which grew out of the tradition of task lists, are better suited to assessing the nature of the job than the quality of performance demonstrated by an incumbent accomplishing a task. Secondly, the checklists which evolved from this tradition include large, often exhaustive, lists of tasks which must be performed for adequate performance but which are cumbersome, at best, for providing feedback or coaching for development. How would a manager convey the feedback on 100-700 task items? A task is done or not done. How can that information from the checklist be helpful for developing an incumbent's
capabilities to accomplish new or different tasks in the future?

A latter day version of the performance checklist is the Behavior Observation Scale (BOS). Latham and Wexley (1977) provide a good description of how items are generated for the BOS. Clusters of descriptive behavioral items are formulated on the basis of data gathered through the critical incident interview technique (p.257). The BOS process is essentially the same process that was used to identify the dimensions and associated behaviors in this study. Each behavioral item becomes a BOS on which the manager/rater records the frequency with which s/he actually observed the employee demonstrate the stated behavior (e.g. 0-19%, 20-39%, 40-59%, 60-79%, 80-100%).

Both the preceding study by Latham and Wexley (1977) and a later study by Latham, Fay and Saari (1979) attempted to identify the optimal empirical approach to developing BOS. Subsequently, proponents of the BARS (BES) methodology (e.g. Bernardin and Kane, 1980) have criticized BOS and argued against their use and interpretation of the percentage ratings for individual behaviors while arguing for BARS (BES) and its Thurstone scaling approach to scaling behaviors as indicators of performance dimensions. As argued previously with respect to BARS, the quality of ratings from either of these
different methods of performance appraisal may rest on the degree to which they facilitate a focus on the dimension related behavior of a job incumbent. The quality of ratings will be effected by the degree to which raters have had a clear sense of the relationship of behaviors to the dimensions being rated and have focused on the relevant, observed behavior of ratees.

Rating theory suggests that conceptual clarification of rating dimensions, in terms of which behaviors to attend to, will help raters as they make assessments of observed performance. The observations of researchers and research reviewers seem to support this proposition. Relying on this body of work as support, this research undertakes to study the consequences of such efforts to clarify the constructs underlying assessment dimensions in terms of their impact on the independence and interrater reliability of dimension ratings.
The Research Hypotheses

The research and theory reviewed to this point suggest two practices which should be associated with high quality ratings when raters observe another person performing on a job and then rate that person on several dimensions considered important for job performance. When rating dimensions are clarified for raters by specifying what kinds of behavior are relevant to each of the different constructs to be assessed (i.e. rating dimensions), ratings should reflect actual performance better than when raters are not so informed. By comparison to the ratings made by uninformed raters, ratings made by raters who actively use the behaviorally clarified rating dimensions to process what they see as they observe the person(s) to be rated should also better reflect performance because the raters focused on dimension-related behavior.

While there is considerable interest in improving the quality of ratings, "quality" has been defined in many different ways. Consequently, many different criteria and methods exist for gauging the quality of ratings (see Saal, Downey, & Lahey, 1980). Rating quality may be reflected in the raters' differentiation of dimensions.
The relative independence of ratings, known as halo, may be indicated by the intercorrelation among various dimension ratings of a person's performance or by the variance among those dimension ratings. Another indicator of rating quality is the level of interrater agreement on ratings for an individual ratee. Halo and interrater reliability were chosen to be the indices used to assess the effects of (1) providing behavioral clarification of rating constructs/dimensions to raters, and (2) of the active use of those behaviorally clarified dimensions during observation of performance.

Research on each of these measures, halo and interrater reliability, suggests a relationship between each of them and the clear articulation and/or use of specific descriptions of the behavior relevant to the rating dimensions. A brief review of the rationale for each choice follows.

1. Differentiation Among Dimensions: HALO

In 1980 Landy and Farr called for a moratorium on studies of rating-scale format, calling in their stead, for more research on the cognitive processes of raters and rating process research. As early as 1920, Thorndike had lamented that raters "are unable to treat an individual as a compound of separate qualities and to assign a magnitude to each of these in independence of the other" (pp. 28-29). This phenomenon, which has become widely known as
"halo," is pervasive. Vance, Winne, and Wright (1983) took up Landy and Farr's challenge and turned from the typical rating-scale-format study of halo to a longitudinal design to look at rater and ratee effects in performance ratings. They concluded that most of the variance in halo indices was attributable to raters and they seconded Landy and Farr's call for more research on the factors affecting raters. This research effort responds to that call.

Ratings of a complex set of performance dimensions are likely to intercorrelate because of rating errors known collectively as "illusory halo," and also because of "true halo,"--i.e., actual interdependence of the assessment dimensions rated. It is not reasonable to expect to be able to identify totally independent aspects of performance for rating. As Smith has noted "independence or orthogonality of dimensions is an ideal goal for statisticians but is unlikely for real behaviors" (Latham et al., 1980). Therefore, clear and specific definitions of the various dimensions are very important in order to equip raters to differentiate appropriately among dimensions (Ghiselli et al., 1981). Rater judgments often reflect attention to factors not intended by the rating instrument. Raters filter perceptions on the basis of implicit schemas or theories (cf. Cooper, 1981). Rater
recall may also be influenced by the biases inherent in judgmental heuristics (Tversky & Kahneman, 1974). Wherry's Theorem 8 (1982; p.534) suggests that use of a checklist of objective cues, such as a list of dimension related behaviors, will reduce the effects of such biases.

In an unpublished study (McBer and Company, 1983) conducted for the U.S. Navy individuals trained to recognize and practice behaviors associated with important job-related skills were asked to make self-ratings on their ability to demonstrate those skills. The mean ratings on the performance dimensions for individuals who had been trained did not differ from the means of the self-ratings by untrained job incumbents. Nonetheless, the ratings of the trained individuals showed more variance across scales (i.e., greater differentiation or less halo) and less variation among the six items on the same scale than did the ratings by the untrained persons. Although it is risky to speculate on the extent of its impact, the behavioral focus and clarification of dimensions for trained persons may have contributed to their ability to differentiate among scales and dimensions of performance. This study is designed to more specifically test the effects of an attempt to clarify rating dimensions.

One fairly common explanation for halo is the
proposition of "insufficient concreteness". The proposition holds that halo can be avoided if raters are provided with rating categories that are highly descriptive, empirically derived, sufficiently specific, and concrete (Cooper, 1981). The behavioral descriptions of dimensions used in this study meet all those criteria. The study itself is intended to provide a test for the foregoing observations about the relationship between halo in ratings and whether raters have seen or used concrete rating-dimension descriptions which specify the kinds of behavior related to each rating-dimension.

2. Interrater Reliability

As Ghiselli, Campbell, and Zedeck (1981) point out "unreliable scores are of little value when we wish to compare two or more individuals on the same measure, to assign individuals to groups or classes, to predict other types of behavior, or to assess effects of various factors on an individual's performance (p. 185)". These are precisely the reasons that the Standards for Assessment Center Operations (1980) specify that each organization should be prepared to demonstrate that its assessors can meet performance standards. Clearly interrater reliability should be among those standards; interrater reliability should be maximized regardless of the assessment technique used by the raters.
What evidence exists for a relationship between behavioral specification of assessment dimensions and interrater reliability? Proponents of the assessment center method contend that when appropriately designed, with a base in job analysis, assessment centers provide a set of content valid assessment opportunities that result in reliable and predictive assessment-dimension ratings. They prescribe the development of dimensions that are based on job analysis and clarified for raters in terms of relevant behaviors (e.g., Byham 1980). As a consequence of careful design, the reported interrater reliabilities for assessment center dimension ratings are fairly good. Sackett and Hakel (1979) reviewed the reliability data from a number of assessment centers; they reported an average interrater reliability coefficient of .69. In a broader review of the assessment center literature Williamson and Schaalman (1980) reported interrater reliabilities ranging from .01 to .93. Unfortunately, they did not report individual coefficients from the studies but rather reported means or ranges for the rs in those studies. Among the rs reported, 15 out of 17 sets of rs included r values in excess of .60 This study is designed to demonstrate that specification of dimension-related behaviors can be shown to relate to interrater reliability of dimension ratings. Measurement theory—in particular
the theory of rating—suggests that such a relationship should exist; the proponents of assessment centers assume it does. Reported interrater reliabilities of assessors actually using behaviorally specific statements of dimensions are quite good. This study intends to demonstrate whether interrater reliability of ratings varies (1) with having access to behaviorally specific descriptions of dimensions or (2) with use of those dimensions by the observer/rater.
In conclusion, the research hypotheses for this study are:

H1: Raters having access to descriptions of rating dimensions which clarify those dimensions in specific behavioral terms prior to observing the behavior of the person they will rate, will make subsequent dimension ratings of the person's observed performance which show greater independence than dimension ratings made by those without access to such descriptions.

H2: Raters having access to descriptions of rating dimensions which clarify those dimensions in specific behavioral terms prior to observing the behavior of the person they will rate, will make subsequent dimension ratings of the person's observed performance on which there is greater interrater agreement than there is agreement on ratings made by those without access to such descriptions.

H3: Raters who use descriptions of rating dimensions which clarify those dimensions in specific behavioral terms as a checklist to record what they observe as they watch a person's behavior, will make subsequent ratings of the person's observed performance which will show greater independence than dimension ratings made by those who have not used such descriptions as checklists while observing the person whose performance is to be rated.

H4: Raters who use descriptions of rating dimensions which clarify those dimensions in specific behavioral terms as a checklist to record what they observe as they watch a person's behavior, will make subsequent ratings of the person's observed performance on which there is greater interrater agreement than there is agreement on dimension ratings made by those who have not used such descriptions as checklists while observing the person whose performance is to be rated.
In some respects these hypotheses may seem to be mundane restatements of the conventional wisdom. Studies abound that compare other rating formats to behaviorally anchored scales. Rating theory, researchers and practitioners all subscribe to the notion that having access to and using clearly stated behavioral descriptions of rating dimensions will result in raters making more reliable, less biased assessments. Sometimes this is a clearly stated assumption. At other times it can be inferred from the predictions made in a research study or the choices made in implementing an assessment process requiring rating. This pervasive nostrum deserves a direct test.

In his theory and theorems Wherry maintains that use of items referring to easily observed behavior categories or use of an easily accessible checklist of such objective cues will help a rater focus and help him/her make more accurate ratings. He talks about memory in general terms and suggests that the preceding methods provide the rater with a constant and objective schema to facilitate accurate recall. In later years, researchers studying the rating process have elaborated his propositions. In their 1980 article Landy and Farr reviewed the related research and proposed a process model of performance rating. They conceived of an actual rating as a linear combination of many sources of
variance, including characteristics of the rater, ratee, rating instrument, and conditions surrounding the rating process. They note, as did Wherry, that the rater brings sets of expectations to the process that may be related to any number of personal and job-related experiences. Any errors in rating, therefore, are not simply properties of the scales or instruments but may result from cognitive differences among the raters. This study was designed to provide different levels of behavioral clarity in the cognitive frameworks that raters bring to rating. It did not try to identify pre-existing cognitive differences, but instead provided a set of cognitive frameworks designed to differ in behavioral specificity. Analyses were used to test for predicted differences in the ratings made by raters who had access to different cognitive frameworks and who used them in different ways.

In June of 1984, DeNisi, Cafferty and Meglino published a cognitive model of performance appraisal. Their model evolved from the social cognition model of category accessibility. Their theory elaborates on Wherry's observation that memory and schema are part
of the rater equation. DeNisi et al. strongly argue that formation of a cognitive representation of behavior is needed to facilitate storage and to enable retrieval of information about the behavior relevant to ratings. Rater judgements are dependent upon recall and integration of relevant information, which in turn are dependent upon prior representation and storage of observations in memory. Prior to making ratings using a standard rating instrument, raters in this study were provided with lists of rating dimensions which were qualitatively different from one another in terms of behavioral specificity. Both the specificity of those lists and the instructions for their use were hypothesized to effect the raters' schemas or preconceived notions and to facilitate categorization, storage and accurate recall of observed performance. Both the levels of specificity and instructions for use can be found in practice, research and theory. While this study does not attempt to map the cognitive processes of the rater per se, it does test the potency of a set of more and less behaviorally specific schemas or clarifications of a sample of dimensions to be used by raters. It also tests for differences resulting from differences in their use.
Most rating research has focused on comparison of specific rating formats, instrument design, specificity of scales or rater training. In this study one simple rating format was used. All raters were given the same amount of time to prepare for rating but no formal training was provided. In these respects the study was different in kind from many studies that preceded it and compared formats, scale specificity or training in terms of the subsequent halo or interrater agreement of ratings. It is not a sophisticated or elaborate study of the cognitive processes of the raters, but it does compare the ratings made by raters who bring different "preconceived notions" (DeNisi et al., 1984) to the rating process. Wherry's theory of rating suggested some means by which raters could be helped to focus on the appropriate schemas for identifying and integrating relevant observations into accurate ratings. Whereas Wherry's theory provides the basis for selection of the specific treatments, DeNisi et al.'s more recent model provides a sound theoretical context, describing the cognitive mechanism which may be effected by attempts at behavioral clarification of the rating constructs.

If the behavioral clarification of constructs has a strong impact on the psychometric properties of ratings, as predicted, then it should be revealed by this study. There will be no variation by condition in either the
rating scale format or the training of raters which could confound the findings. A positive set of findings would support Wherry's theorems and could turn the conventional wisdom from nostrum to research based prescription. A negative set of findings should put researchers and practitioners on guard not to assume that behavioral clarification of rating dimensions can or will significantly improve multidimensional assessments made using rating scales.
This research study was not based in the theory and research surrounding a single, human resource management practice. In a wide range of organizational contexts managers assess individuals in terms of dimensions which reflect the behaviors demonstrated by those who perform the job successfully. Performance reviews, candidate interviews, and assessment centers all involve raters in observing, categorizing, and evaluating the behavior demonstrated by another person. Consequently a variety of operational choices were available to the selection of methods to implement and test the preceding research hypotheses.

This section addresses each of the major operational choices in turn, briefly explaining the option chosen and the rationale for that choice.

The Raters

Four groups of 30 raters each were required as subjects for this study. The subject population was chosen so that they would (overall) have had only limited experience in evaluating others, and no specific experience with the rating dimensions to be used in the study. These characteristics were desired so that subjects did not bring conflicting or confounding
experience to bear on the experimental tasks.

Navy officers attending Officer Indoctrination School (OIS) were selected because they would have had limited experience and limited authority to evaluate others' performance prior to their upcoming first tours as officers. More importantly, they would have a real impending need to rate the performance of subordinates. This was seen as desirable because it would increase the actual relevance of the experimental task to them.

The Treatments

The hypotheses imply 3 conditions. A fourth was added at the request of a committee member in order to address a related question not specifically raised in the hypotheses. The four treatments and the rationale for their design is described below.

- Group 1 received a list of rating dimension names only; no clarification of the constructs represented by those names was provided. They were told that after observing people performing job tasks they would be given rating forms to record
their evaluation of those people in terms of the list of dimensions provided. The rating forms were labelled with dimension names only.

This was the minimum treatment. It was designed this way because a great many organizations use rating forms of this kind to make assessments, on the assumption that the adjective labels or trait names are sufficient to indicate to the rater what s/he is to assess following observation of a ratee's performance.

Group 2 received a list of behaviors that were associated with each of the dimensions to be rated. They were also instructed that they would be asked to rate the individuals and their performance in terms of the dimensions on the list provided. They received the same rating forms as Group 1.

This treatment was chosen because it epitomized the state of the practice. Many organizations now label assessment forms or give assessors instructions that indicate which behaviors are relevant to the dimensions to be rated, but few organizations bother to scale and rate behaviors. Consultants and researchers usually recommend providing behavioral definitions of performance dimensions as a minimum condition for introducing existing performance-rating
dimensions to new raters or new rating dimensions to experienced raters.

- Group 3 was asked to use the lists of dimension-related behaviors that they (and Group 2) received, as a checklist for recording observed behavior as they watched the ratees' job performance. They were told that they would later be asked to rate the individuals in terms of the dimensions on their checklists.

As indicated previously, several researchers have suggested that this approach has considerable potential for reducing halo. In practice, a number of organizations have begun to use a behavior-scoring approach (similar to use of a behavior checklist) to observe and assess performance systematically. This treatment was chosen as a means of inducing maximum attention to and use of the behaviors that provide evidence of the constructs to be rated.

- Group 4 simply used the lists of dimension-related behaviors as checklists, and was not told to make ratings. The group was included in the study for comparison with Group 3--to provide a means to determine whether raters attended to behavior and used the checklist differently when they
anticipated having to rate the behavior observed.

In his theory of rating Wherry predicted that ratings gathered under research conditions would be more accurate than ratings gathered under administrative conditions. In reviewing the research literature related to this prediction Landy and Farr (1980) concluded that administrative ratings tend to be more lenient, but not necessarily more or less reliable and valid than ratings gathered for research purposes. All the ratings and scoring data collected from Ss in this study were used for research purposes; the Ss all knew this.

The broader implication of the preceding research, however, is that the rating or scoring done by Ss may be effected by their anticipation of how their recorded judgements will be used. Ss in Condition 3 were asked to use the behavior list as a checklist of their observations; the checklist was divided into ten groups of behavior corresponding to dimensions of performance on which they would later be required to evaluate the two persons they observed. Ss in both conditions 2 and 3 had access to the same detailed lists of
dimensions and their associated behaviors. If ratings from raters in Condition 3 were more reliable and had less halo than ratings from Condition 2 then one conclusion would be that the superior Condition 3 ratings were associated with the treatment effects of actually using the behavior list as a checklist.

This conclusion assumes that the Ss used the checklist and subsequently made ratings reflecting what they observed and checked. It would be possible, however, for Ss to have attended to observed behavior and to have used the checklist to reflect their preconceived notions of how they would rate the persons. (i.e. OBSERVE + CHECK \(\Rightarrow\) RATING vs. ANTICIPATED RATING \(\Rightarrow\) OBSERVE + CHECK.) In order to eliminate the latter explanation Condition 4 was included. The investigator compared the checklist scoring done by Ss who were merely told to record what they observed (Condition 4) and those who were told to do so because they would later be asked to rate the persons they observed (Condition 3). A finding of no difference between checklist scores from Ss in conditions 3 and 4 would be needed to demonstrate
that the raters observations were not effected by their anticipation of how their judgements would be used, and thus to strengthen the interpretation that the differences in ratings from conditions 2 and 3 were treatment effects.

The Stimulus

Most of the multidimensional assessment ratings of individual performance that are made in organizations are based on observation of complex human interactions. This is true for the standard annual performance review, and it is also true of assessment center ratings. Consequently, rating observation of a single individual, doing a single well-defined task, was rejected as too simple a rating task with limited potential for generalization of findings. This study, therefore, used as a stimulus a film of several individuals performing a job that was familiar to and/or of interest to the subjects.

The 1949 film Twelve O'Clock High was chosen as the stimulus for this study for several reasons. Films of individuals in actual assessment situations (such as group exercises in assessment centers) were considered as possibilities. After reviewing a number of videotapes from several different assessment centers, the investigator eliminated that kind of film as a poor alternative. The available videotapes tend to be records of exercises in which people talk, argue, and plan on
the basis of materials which the viewer never sees; little or no active behavior is shown. Also, even the best quality videotapes of assessment exercises tend to have only fair to poor audio and video quality,—especially by comparison to commercial films. Twelve O'Clock High, by contrast, contains a clear record of individuals actively involved in a variety of job tasks and interactions. By comparison to assessment exercises on videotape, this film is both more engaging and more directly analogous to the kinds of observation opportunities a rater/supervisor usually has to observe others prior to rating their behavior. These characteristics of the film are especially important for involving the raters with the task given the neccessary truncation of observation time available to raters in experimental situations (such as this one). The film is also of interest to Navy audiences and has credibility with them because of the military scenario and organizational structures conveyed. They can identify with the characters and get involved with them in a way that would be unlikely if they were to be asked to observe and rate civilian managers in an assessment exercise. Lastly, the film is in the public domain and, therefore, available for use in further studies by other researchers.
The Rating Task

The raters were asked to make ratings of 10 performance dimensions for 2 ratees each. This combination was chosen for the following reasons. Ten dimensions were rated in order to model a realistic assessment task, generalizable to assessment centers and performance appraisal practices. Moreover, a fairly large number of dimensions was necessary in order to have a large enough number of judgments to get fairly stable estimates of interrater reliability and of halo. Another approach to getting a large enough number of judgments could have been to increase the number of ratees and reduce the number of ratings made for each ratee; but this was not a reasonable option for this study given the hypothesis that halo and independence of ratings will vary with the treatment conditions. It was not empirically sound to draw conclusions about the complex intercorrelation of rating dimensions based on ratings of a limited number of dimensions.

Two ratees were assessed, not one. This choice was made to allow the researcher to look at ratee differences as a post-hoc analysis if desired (although, clearly, neither the research study as a whole nor the two-ratee design was intended to provide a test of ratee effects). But more importantly from the researcher's perspective, a two-ratee design provided a more reasonable approximation
of organizational realities than did a one-ratee design, improving the basis for interpretive generalizations from the findings.

The Dependent Measures: Halo

A variety of possible indices exist for the measurement of halo—the lack of the desired independence of ratings.

One measure of the dependent variable, halo, is the standard deviation of ratings for each ratee. This is a measure of intraratee variance across rating dimensions. It measures co-occurrence (ratings at the same level) across dimensions such that perfectly haloed ratings have intraratee variance equal to 0. The variance method reflects a more restrictive view of halo than some other measurement approaches (Cooper, 1981; p.219). In this study the standard deviations of ratings were used as measures of the dependent variable in an analysis of variance which compared the level of halo in sets of ratings from the different conditions.

Analyses were conducted that assessed the level of halo in dimension ratings by examining the intercorrelation among different dimension ratings across raters. This operational definition of halo was chosen because calculating interdimension correlations provides a covariance measure that equates perfect halo with
correlations equal to one. If dimensions differ in their mean levels, distributions, or social desirability, perfect covariance can exist despite intercategory variance. Since the mean, distribution, or desirability of rating dimensions may vary, this was chosen as a second means to assess halo of ratings.

The Dependent Measure: Interrater Reliability

The degree to which raters agree on ratings was assessed by calculating the intercorrelation of each rater's ratings with the ratings made by each of the other raters in the treatment group. The mean of these intercorrelations was calculated as a summary measure of the interrater reliability within each treatment group. The summary measures were then used for comparing the interrater reliabilities of the different treatment groups.

Optional Supplementary Hypothesis: Operational Definitions

Implicit in the two primary hypotheses are two premises:

(1) Raters operating from clear, behavioral definitions of assessment dimensions will make better ratings in terms of independence of ratings (or lack of halo) and interrater reliability.

(2) Clarity of the operational definitions actually used by raters will vary by treatment.

Dimension definitions, written by members of the different
treatment groups after they have made ratings, were used
to check these assumptions. (They were asked to write
definitions for a sample of three dimensions.)

Hence, a possible supplementary hypothesis:

HSl: The criteria which raters report taking into
consideration when rating an individual's
performance will vary in content (e.g.,
specificity, behavioral language, distinctive
character) depending upon the treatment group to
which the raters belong.

If the predicted treatment effects were not found, then
the definitions could be content-analyzed to test
Hypothesis Sl. This could be considered a manipulation
check. The analysis could only be suggestive, however,
since no definitions could be collected prior to the
treatment without risk of sensitizing subjects.

Optional Supplementary Hypothesis: Scoring

Group 4 was asked simply to score observed behaviors
in terms of the checklists of dimension related behaviors
provided to the group members. This treatment was
included because of an implicit assumption in the design:
the way in which raters rate will be affected by their use
of the checklist to record observed behaviors. It could
be that individuals who expected to be called on to rate
anticipated rating and used the checklists to support the
ratings they expected to make. In other words, rather than
the ratings reflecting the observed behavior that was
checked, the checked behavior reflected the ratings. If anticipation of rating did not in some way affect checklist use, then checklist scores for Groups 3 and 4 should not significantly differ.

Hence, an additional supplementary hypothesis:

HS2: Individuals who know they will be asked to make dimension ratings will score observed behavior using checklists in the same way as those who are asked only to score behavior.

The scores from Groups 3 and 4 can be compared in order to test this hypothesis.

Raters in conditions 1, 2 and 3 will all make judgments on the degree to which each ratee demonstrates each of the dimensions through his performance. Raters in Condition 3 will make those ratings after using the checklist to record when they observe behaviors relevant to the rating dimensions. Since Ss in Condition 4 are instructed to use the same checklist, in the same way, but make no ratings, their checklist scores will be compared to those from Condition 3. No comparison of Condition 4 with Condition 1 or 2 will be possible since checklist scores and the process by which they are generated are not comparable to ratings or the rating process. (See Appendix A for definitions which distinguish scoring from rating [Bales et al., 1979].)
METHOD

Subjects

The subjects were 119 new Navy officers assigned to OIS in Newport, Rhode Island. They were all college graduates, most of them coming directly to OIS after graduating from school. They were in the first week of an 8 week program to learn procedures and to learn basic leadership and management skills. During the first week in the OIS program they were still focusing on operating procedures and military systems. They had not, as yet, attended any management classes which could in any way have affected their responses to the experimental tasks. Subjects were randomly assigned to conditions. Data were collected from 2 different classes (June 1984, July 1984), but at the same point in their programs (i.e. 1 week into the schedule). Demographic data on the subjects, by condition, can be found in the Results section (Table 7).

Stimulus

All the subjects viewed one hour and 10 minutes of the film Twelve O'Clock High (Los Angeles, Calif.: Twentieth Century Fox Film Company, 1949). It was shown on a big screen in a large classroom. Subjects were specifically asked to attend to the behavior of two officers (Davenport and Savage). The two characters each perform the job of
commanding officer during this portion of the film. Each interacts with his subordinates, and with his superiors; each attends to tasks required by the job. A full synopsis of the portion of the film shown can be found in Appendix B.

Selection of rating dimensions

A pilot study was conducted in order to test and refine the materials to be used in the later data gathering. The final 10 dimensions were selected on the basis of the pilot data. The dimensions sought would be a group of constructs which would all seem socially desirable and relevant to the commanding officer job as portrayed in the film. The investigator made this judgment. At the same time, some real variance was needed between the scores for the two officers and among the scores for any one officer. The pilot data provided a small sample of information on these counts.

A variety of different constructs and associated behavior were considered for use as rating dimensions. All of these dimensions were derived by means of a variety of job analysis known as job competence assessment. Job competence assessment is a person-focused (as opposed to task-focused) method of job analysis which uses a combination of critical incident interview data and
content coding to identify patterns in how incumbents approach their jobs; the data are then analyzed to identify which of those behavior patterns characterizes the outstanding performers among a set of job incumbents. Those characteristics which distinguish outstanding performers are called competencies and can be used as criteria or dimensions for assessing the performance of other job incumbents or job candidates.

The specific groups of behaviors which become examples of a dimension of performance are identified in a series of steps. First, a thematic analysis of all the critical incident interviews is performed. This is accomplished by recording behaviors that each job incumbent performs which seem to be associated with his/her success in dealing with some aspect of the job. A team of analysts reviews this volume of data to identify groups of behaviors which recur in a number of incidents, from a number of interviewees, doing one or more of the tasks required for performance of the job. Since the desired product of this process is a list of dimensions related to superior performance, the focus is kept on behaviors in situations which are either high points or low points on the job; special attention is focused on identifying behaviors which are demonstrated by a criterion identified group of superior performers, but are lacking or
infrequently exhibited by average performers.

The behaviors identified in this process are grouped together on a conceptual basis. The behaviors may reflect the job incumbents' use of specific knowledge, inherent abilities, learned skills or motives for performing the tasks on the job. All are considered to be examples of a person's characteristic manner of performing the job. They are grouped conceptually and labelled with phrases that capture their conceptual link and describe an aspect of how the job incumbent acts on the job. Since this is a person-focused job analysis process, the resulting list of dimensions looks like a list of characteristics of a person. They are not to be confused with seemingly similar lists used by some organizations—these are not dimensions of personality. Personality dimensions are individual, sometimes immutable, predispositions to react to situations in a particular way. The dimension list described here characterizes how successful job incumbents may be recognized in terms of behaviorally specific examples of how they may behave in the job context. The competency or dimension names serve as convenient summary labels referring to related behaviors in the individual's job performance.
Many organizations use a list of dimensions for evaluation of individuals and their job performance. The list may look very much like this experimental set of dimensions. Since the list may look like a list of personality variables, the dimensions may be construed by raters to reflect either the job incumbents abilities, or personalities, or aspects of their job performance. In this study, that ambiguity of focus and intent is a hypothesized contributor to halo and low interrater reliability. In those conditions where behavioral examples of the dimensions are provided to the raters the focus is more appropriately directed to aspects of an individual's demonstrated job performance.

The dimensions under consideration came from several different job competence studies. Thirteen dimensions were derived from a study of naval senior officers in executive officer and commanding officer jobs. An assortment of other competencies were selected as possibilities from other studies on the following bases.

(1) They represented socially desirable constructs which were related to superior performance in some circumstances.

(2) At least one of the two officers to be rated demonstrated the competency during the portion of 12 O'Clock High to be shown.

Fourteen dimensions were chosen for use in the pilot. They are listed in Appendix C.
The pilot study sample consisted of 30 Ss; 10 each in conditions 1, 2, 3. They performed the same tasks as the final study population did except that they were asked to observe and then rate the ratees in terms of fourteen characteristics considered important for performance.

The data were reviewed and a set of 10 dimensions chosen such that each ratee showed some variance across dimensions and the two ratees differed from one another. (See Appendix D for the pilot rating data reviewed to make these determinations.)

The final dimensions chosen were then randomly ordered so that, for example, the dimensions taken from the original senior officer competencies did not all appear together and the dimensions on which each officer excelled were randomly distributed. The final list and order are shown below; detailed behavioral descriptions of the dimensions can be found in Appendix E as part of the materials given to Ss in condition 2.

**RATING DIMENSIONS**

Informed Judgment  
Command Influence  
Building Morale  
Monitoring for Results  
Empathy  
Conscientious Use of Discipline  
Planning  
Interpersonal Diagnosis  
Sense of Responsibility  
Initiative
Treatment

Overall Procedure

Each of the 119 subjects received a large manilla envelope full of materials. They were asked to not open their envelopes until all participants had entered the room, received their materials and been seated. The Ss from each condition sat together in the section of the room designated for participants with particular numbers marked on their envelopes; thus participants in condition 1 had a number between 100 and 130 on their envelopes and sat together in the section marked 100s, Ss from condition 2 sat with other 200s, etc. (This seating arrangement was intended to help keep Ss focused on their own material and avoid looking at, comparing, or otherwise being distracted by the materials and activity of a person in an adjacent seat who might be in a different condition.)

Subjects were given a brief introduction to the investigator and the task (see appendix for instructions scripts) and were then asked to remove materials from their envelopes. Each envelope contained 3 sets of materials, each printed on a different color paper. Thus the investigator could specifically request that only certain materials be removed at any one time. All
subjects were asked to remove their assessment dimension materials only (i.e. those clipped to a yellow instruction sheet),-- these of course differed by condition. The investigator then read aloud instructions to review the materials and to think about what the dimensions mean. They were told that they would later be using the dimensions on their lists to evaluate the performance of two people who they would have a chance to observe in the upcoming film. They were given 15 minutes to review the dimensions. Each participant had a copy of the instructions read by the experimenter and also specific instructions for his/her condition printed on the dimension sheet. (Copies of all instructions are included with the sample materials in the Appendix.)

After 15 minutes the film was introduced by the investigator. It ran for one hour and 10 minutes. Participants were asked not to speak to each other during the film.

After the film was stopped Ss were asked to take the white pages from their envelopes. For conditions 1, 2, 3 these were the rating scales. Condition 4 removed a white sheet from their envelopes which explained that the others were performing a task that they were not required to do, and that in about 5 minutes they would all have one more
task to do. Group 4 was asked to wait quietly.

Once the Ss in conditions 1, 2 & 3 had completed their ratings, all Ss were instructed to remove the final questionnaire from their envelopes and complete it. This was the data sheet for gathering demographic information and information to use as a manipulation check.

All materials were collected when complete.

Materials
Copies of all materials are included in Appendix E.

Dimension lists. The dimension lists were presented in different formats, with different levels of detail depending upon condition.

- Condition 1 received a list of ten dimension names. The list did not include any examples, description, or definitions of the dimensions.

- Condition 2 received a list of dimensions which included descriptions of the kinds of behavior which demonstrate each of the dimensions. The information was presented on one oversize page so it could all be looked at at once.

- Condition 3 had the same information as Condition 2. The list for this condition also had two columns of spaces where the Ss were to check off behavioral evidence of a dimension when they saw a ratee demonstrate it. All the information was presented on one oversize sheet.

- Condition 4 received a dimension list identical to that for Condition 3 except that the instructions for Condition 4 did not mention that they would have to make ratings after using the checklist.

Rating scales. The rating scales used by Ss in conditions 1, 2, 3 were in the format of a 5-point Likert scale. The
format was borrowed from the standard assessor materials provided by a major assessment center supplier. For each of the 10 dimensions the Ss were asked to circle the number which corresponded to their evaluations of the individual's performance in terms of that rating dimension. Ss made ten ratings for Savage and ten for Davenport. The scale was anchored with adverbial phrases indicating the degree to which the characteristic was shown. The scale is shown below.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>quite a lot shown</td>
<td>a moderate amount shown</td>
<td>a small amount shown</td>
<td>very little or none shown</td>
</tr>
<tr>
<td>shown</td>
<td>shown</td>
<td>shown</td>
<td>shown</td>
<td>shown</td>
</tr>
</tbody>
</table>

Participant questionnaires. This supplementary set of questions was completed by all participants. It was used to gather information about their prior experience, and feelings about the materials they had used as well as to tap their perceptions of what the different dimensions meant. The questionnaire was 4 pages in length and took about 10 minutes to complete.
ANALYSES

The dimension ratings collected from Groups 1, 2 and 3 were analyzed as tests of Hypotheses 1, 2, 3 and 4. Analyses of variance and subsequently correlational analyses were used to examine the data relevant to the hypotheses predicting effects on halo. Correlational analyses were used for the major investigations of the reliability data. Supplementary analyses were added to further explore the data.

Halo: Analysis of variance

One way of looking at halo is to look at the variation across dimension ratings of a ratee made by one rater. This approach treats the dimensions as alternate measures of the person rated. By looking at the standard deviations of a rater's ratings in this way one assesses first the extent to which each rater within a treatment differentiates among dimensions rather than the overall level of differentiation among dimensions by all raters within a condition.

In order to test the differences in halo by condition an analysis of variance was performed. The dependent variable was the standard deviation of each rater's ratings of one ratee. Since each S rated two
ratees there were 60 standard deviations per condition (30 raters x 2 ratees) on which to base these analyses. Subsequently, the means of the standard deviations for the different groups of ratees were used to compare pairs of conditions in terms of halo.

Halo: Correlational Analyses

Hypothesis 1 predicted that the independence of dimension ratings would vary by treatment. A more common way of referring to the independence of dimension ratings is to talk about the "halo" that appears in ratings, indicating a lack of independence and a potentially spurious relationship among those ratings. These analyses of halo in the rating data used a correlational measure as the dependent variable representing the degree of independence among a set of dimension ratings given to a ratee's performance.

For each dimension there were 60 ratings made by raters in each condition of this design (i.e., 30 raters x 2
ratees). The operational definition of halo in use in this analysis was "the intercorrelation among different dimension ratings using ratees' scores (over raters) for each dimension as data points" (Saal, Downey & Lahey, 1980). The analyses, therefore, consisted of correlating ratings on each dimension with those for each of the other dimensions in turn. This process produced for each treatment group, a matrix of the pairwise correlations of all dimensions. The mean of the correlations from the matrix for each treatment group's dimension ratings was the summary measure that was used to compare the level of halo by treatment groups. Starting with r to Z transformations, standardized z scores for the conditions were then used to compare the mean of the correlations for Condition 1 with the mean for Condition 2, Condition 2 with Condition 3, and Condition 3 with 1.

**Interrater Reliability: Correlational Analysis**

Hypotheses 2 and 4 predicted that interrater reliability would vary by treatment. This prediction was tested by correlating each rater's ratings with the ratings made by each of the other raters in the treatment group. Each rater made 20 ratings (i.e., 10 dimensions x 2 ratees). The analysis of these data consisted of
correlating each rater's ratings with each of the other rater's ratings, in turn, to produce a matrix of the pairwise correlations of all the raters' ratings. The mean of the correlations from the matrices for each treatment group was the summary measure of interrater reliability (Guion, 1965) used to compare the treatment groups. Starting with r to Z transformations, standardized z scores for the mean r values were then used to compare the means for Conditions 1 and 2, Conditions 2 and 3, and Conditions 1 and 3.
RESULTS

Hypothesis 1

The test of Hypothesis 1 was conducted using analysis of variance. Halo in ratings was operationalized as the standard deviation of ratings within rater for each ratee. Thus two measures of halo were calculated for each rater within condition and these values were used as the dependent variables for the ANOVAs. Each was a one way ANOVA of raters by condition with halo as the dependent variable. Although analysis of variance is a robust analytical method, it is most appropriately used with dependent variables which meet the assumption of normal distribution within the population. Distribution information is presented in Table 1 below. The standard deviation of ratings are very nearly normally distributed. Indices of skewness and kurtosis are included with the distribution information in order to provide a quick sense of how the dependent variable is distributed. (No further interpretation of the differences between these indices for the three conditions should be ventured, however, since it is seldom advisable to compute these measures when N is less than 100 [McNemar, 1969; p. 27].)
TABLE 1

Distribution of the dependent variable for ANOVA:
Standard deviation of ratings

<table>
<thead>
<tr>
<th></th>
<th>MEAN</th>
<th>MEDIAN</th>
<th>MODE</th>
<th>RANGE</th>
<th>SKEW</th>
<th>KURTOSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL N=180</td>
<td>1.17</td>
<td>1.20</td>
<td>1.20</td>
<td>1.84</td>
<td>-.87</td>
<td>1.29</td>
</tr>
<tr>
<td>CONDITION 1 n=60</td>
<td>1.08</td>
<td>1.08</td>
<td>1.08</td>
<td>1.70</td>
<td>-.74</td>
<td>.81</td>
</tr>
<tr>
<td>CONDITION 2 n=60</td>
<td>1.17</td>
<td>1.23</td>
<td>.42</td>
<td>1.84</td>
<td>-.92</td>
<td>.99</td>
</tr>
<tr>
<td>CONDITION 3 n=60</td>
<td>1.27</td>
<td>1.25</td>
<td>.99</td>
<td>1.09</td>
<td>-.01</td>
<td>-.23</td>
</tr>
</tbody>
</table>
Summary ANOVA tables for the halo analyses follow.

### TABLE 2

ANOVA for HALO defined as s.d.of ratings: Conditions 1, 2, 3

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups:</td>
<td>2</td>
<td>1.0631</td>
<td>.5315</td>
<td>4.808</td>
<td>.0093</td>
</tr>
<tr>
<td>CONDITION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups:</td>
<td>177</td>
<td>19.5682</td>
<td>.1106</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>20.6316</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 3

A COMPARISON OF MEANS FOR THE DV: S.D.OF RATINGS

<table>
<thead>
<tr>
<th>CONDITION</th>
<th>MEAN</th>
<th>95% CONFIDENCE INTERVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.08</td>
<td>.98-1.18</td>
</tr>
<tr>
<td>2</td>
<td>1.17</td>
<td>1.07-1.26</td>
</tr>
<tr>
<td>3</td>
<td>1.27</td>
<td>1.20-1.33</td>
</tr>
</tbody>
</table>
The ANOVA reveals a significant treatment effect in the dependent variable halo as defined in these analyses. Table 2 presents a comparison of halo across all 3 conditions. The F test for this fixed effect ANOVA provides a test of a linear hypothesis model. The significant F indicates that halo decreases across conditions as predicted.

Hypothesis 1 predicted that raters who had access to descriptions of rating dimensions which clarified those dimensions in specific behavioral terms would produce ratings with less halo than would raters who had not had access to those detailed behavioral descriptions of the rating dimensions. Since raters in conditions 2 and 3 each had access to such lists they would each be expected to produce ratings with less halo than would raters in Condition 1.

Comparisons of pairs of conditions can be made by reviewing the data in Table 3. The mean level of halo (measured by standard deviation) in ratings from Condition 2 does not significantly differ from that for Condition 1. (The difference does not meet the criterion for a .05 significance level, but would meet the marginal criterion of alpha equal to .10.) Thus, providing raters with a specific listing of dimension-related behaviors does not appear to be significantly related to lower levels of halo. The difference in the mean level of halo (s.d of
ratings) from conditions 1 and 3 is significant. Thus, raters who have access to the lists of behaviors related to dimensions and use them as checklists demonstrate significantly lower levels of halo in their ratings than do those raters who don't even have access to such lists.

Correlational analyses were performed on ratings from a total of ninety (90) raters, thirty (30) in each of the three conditions. Each S made two sets of 10 ratings, one set for each of the two ratees. (A table of the mean dimension scores for each ratee across conditions can be found in Appendix F.) Rating scores on each of the 10 rating dimensions were correlated with the ratings on each of the other dimensions in turn. This 10x10 correlation matrix was based on 60 cases (30 Ss x 2 ratees). One matrix was produced for each condition. These matrices may be found in Appendix G. The mean intercorrelation for each matrix was calculated and they are shown in the following table.

TABLE 4

Hypothesis 1: The effects of behavioral clarification of dimensions on the subsequent HALO among dimension ratings

Mean interdimension correlation

<table>
<thead>
<tr>
<th>CONDITION 1 n=60</th>
<th>CONDITION 2 n=60</th>
<th>CONDITION 3 n=60</th>
</tr>
</thead>
<tbody>
<tr>
<td>.30</td>
<td>.18</td>
<td>.17</td>
</tr>
</tbody>
</table>
A comparison of mean r's using an r to Z transformation and comparison of the standardized z to the normal probability curve revealed that these differences in r were not statistically significant.

**Hypothesis 2**

Hypothesis 2 predicted that raters who had access to descriptions of rating dimensions which clarified those dimensions in specific behavioral terms would produce ratings on which they agreed with one another better than would raters who had not had access to such descriptions of the rating dimensions. The levels of interrater agreement on ratings were calculated for raters in each condition. Since raters in conditions 2 and 3 each had access to lists describing the behavior related to each dimension prior to observation and rating, but Ss in Condition 1 never saw such descriptions, the level of interrater agreement on ratings from conditions 2 and 3 were each compared to the level of agreement among raters in Condition 1.

For each of the 30 raters in a condition, his/her 20 ratings were correlated with the 20
ratings made by each other rater in turn. In this way a 30x30 matrix of interrater correlations was calculated for each condition. The mean intercorrelation was calculated for each matrix. The results are shown in the following table.

**TABLE 5**

Hypothesis 2: The effects of behavioral clarification of dimensions on subsequent interrater reliability of dimension ratings

<table>
<thead>
<tr>
<th>CONDITION</th>
<th>MEAN r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (n=30)</td>
<td>.57</td>
</tr>
<tr>
<td>2 (n=30)</td>
<td>.44</td>
</tr>
<tr>
<td>3 (n=30)</td>
<td>.56</td>
</tr>
</tbody>
</table>

Note: r is based on 30 raters in each condition making 20 ratings each.

In contradiction to the prediction, interrater reliability of dimension ratings does not increase across conditions.

In fact, the preceding correlations reveal a pattern in which raters in Condition 2 who were provided with detailed behavioral descriptions of dimensions show a decrease in interrater reliability; no such decrease is apparent when the raters get and use those behavioral descriptions. A comparison of r's using
the r to Z transformation and then comparing the standard z scores did not reveal any significant differences among the rs shown in the preceding table.

**Hypothesis 3**

Hypothesis 3 predicted that raters who actually used the behavioral descriptions of dimensions as checklists to record what they observed would show less halo in their ratings than would raters who didn't use the descriptions of dimensions in this way.

Table 2 presents the analysis of variance which was conducted as a means of assessing halo effects. Table 3 presents data showing a significant treatment effect when the mean levels of halo (s.d. of ratings) for conditions 1 and 3, and 2 and 3, were compared. Actual use of the dimension related behavior lists as checklists was associated with levels of halo lower than the halo among ratings from raters without access to such lists. Their level of halo is also lower than the halo in ratings made by raters who had the lists but did not use them as checklists for recording observed behavior.

The interdimension correlations of ratings made by those raters who used the dimension definitions as checklists (Condition 3) were compared to the correlations from raters in the other conditions. The data are presented in Table 5. None of the differences among the conditions was significant.
Hypothesis 4

Hypothesis 4 predicted that raters who actually used the behavioral descriptions of dimensions as checklists to record what they observed would show greater interrater reliability among their ratings than would raters in other conditions. The interrater correlations of ratings for each condition are presented in Table 5.

The figures in Table 5 reveal a pattern in which raters using the checklist approach in Condition 3 seem to agree among themselves as much as do raters in Condition 1 and more than raters in Condition 2. However, as indicated previously, r to Z transformations and subsequent comparison of standardized z scores revealed no significant differences among these levels of interrater reliability.

Supporting Analyses

Demographic data and participant responses to questions on the final participant questionnaire were analyzed in order to assess whether the preceding lack of significant results could be attributed to either differences among participants by condition or to a failure to affect their perceptions of the dimensions and their meanings.
Table 6 provides demographic profiles of the participant population—overall and by condition. Since Ss were randomly assigned to condition the investigator expected some variation in the profiles of the groups but no systematic or large variation that could be related to differential rater responses by condition.

**TABLE 6**

Demographic profile of Ss by condition

<table>
<thead>
<tr>
<th>CONDITION</th>
<th>DEMOGRAPHIC VARIABLE</th>
<th>1 (n=30)</th>
<th>2 (n=30)</th>
<th>3 (n=30)</th>
<th>4 (n=29)</th>
<th>OVERALL (N=119)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEX -female</td>
<td>23</td>
<td>23</td>
<td>27</td>
<td>31</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>-male</td>
<td>77</td>
<td>70</td>
<td>73</td>
<td>66</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>CORPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Medical</td>
<td>33</td>
<td>13</td>
<td>40</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>-Dental</td>
<td>10</td>
<td>17</td>
<td>13</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>-Judge Advocate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>General</td>
<td>37</td>
<td>37</td>
<td>23</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>-Nurse</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>-Nuclear</td>
<td>17</td>
<td>17</td>
<td>3</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>-Civil Engineering</td>
<td>--</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>PRIOR SERVICE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-yes</td>
<td>40</td>
<td>27</td>
<td>37</td>
<td>31</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>-no</td>
<td>60</td>
<td>67</td>
<td>63</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>HELD SUPERVISORY JOB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-yes</td>
<td>67</td>
<td>63</td>
<td>76</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>-no</td>
<td>33</td>
<td>30</td>
<td>23</td>
<td>38</td>
<td>33</td>
</tr>
</tbody>
</table>

*NOTE: All numbers in the table represent percentages within the condition. When percentages do not add to 100 some Ss did not respond to the question.
Overall, one third of the study participants had prior service. Typically this means that they had served as enlisted and had either completed a college degree while serving or had left the service, and after going to college, were now re-entering as officers. Also, note that two-thirds of the participants claim to have held a supervisory job in the past. As indicated earlier, this population was chosen because they would presumably have had limited experience and authority to evaluate others and because they would have had no experience with these rating dimensions. The demographic data indicate a higher level of experience in dealing with subordinates than had been expected, nonetheless, it is unclear from this data whether the Ss had much rating experience. Since they did not have experience with these dimensions and by virtue of their supervisory experience they presumably have some understanding of the variability of performance, this sample still seems to fit the needs of this investigation.

**Manipulation check**

Ss were asked to respond to questions about their perceptions of the materials they had been given, their experience, and their feelings of preparedness. (See Appendix E for a copy of the participant questionnaire; the following discussion refers to the first five
questions.) Responses were given on a 5-point Likert format scale. Higher numbers reflected more positive responses; the responses would be expected to increase across conditions in order to be in synchrony with the predicted effects of the experimental treatments. (Responses to the question concerning level of experience would have to be approximately equal across conditions in order to not confound any treatment effects.) The mean responses to those questions and comparisons of the condition are presented in Tables 7 through 13.
TABLE 7

Responses to the manipulation check questions:
Mean responses by condition

<table>
<thead>
<tr>
<th>QUESTION</th>
<th>CONDITION 1</th>
<th>CONDITION 2</th>
<th>CONDITION 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>..experience</td>
<td>2.8</td>
<td>2.7</td>
<td>3.0</td>
</tr>
<tr>
<td>..list helped</td>
<td>3.3</td>
<td>4.1</td>
<td>3.9</td>
</tr>
<tr>
<td>..felt</td>
<td>2.6</td>
<td>3.0</td>
<td>3.1</td>
</tr>
<tr>
<td>prepared</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>..knew what</td>
<td>3.4</td>
<td>3.7</td>
<td>3.6</td>
</tr>
<tr>
<td>dimension the behavior</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>related to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>..seemed to be</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>different characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 8

Responses to the manipulation check:
Condition comparisons by t-test

<table>
<thead>
<tr>
<th>QUESTION</th>
<th>1 vs 2</th>
<th>2 vs 3</th>
<th>3 vs 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>..experience</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>..list helped</td>
<td>***</td>
<td>--</td>
<td>**</td>
</tr>
<tr>
<td>..felt</td>
<td>*</td>
<td>--</td>
<td>**</td>
</tr>
<tr>
<td>prepared</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>..knew what</td>
<td>*</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>dimension the behavior</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>related to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>..seemed to be</td>
<td>*</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>different characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*** prob. less than .001
** prob. less than .05
* prob. less than .10
### TABLE 9

**ANOVA for RATER EXPERIENCE**

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups:</td>
<td>2</td>
<td>1.2667</td>
<td>0.6334</td>
<td>0.713</td>
<td>0.4928</td>
</tr>
<tr>
<td>CONDITION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups:</td>
<td>87</td>
<td>77.2333</td>
<td>0.8877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>78.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 10

**ANOVA for PERCEPTION THAT LIST HELPED**

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups:</td>
<td>2</td>
<td>10.4000</td>
<td>5.2000</td>
<td>6.283</td>
<td>0.0028</td>
</tr>
<tr>
<td>CONDITION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups:</td>
<td>87</td>
<td>71.9999</td>
<td>0.8276</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>88</td>
<td>82.3999</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 11

ANOVA for FEELING PREPARED

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups: CONDITION</td>
<td>2</td>
<td>3.2668</td>
<td>1.6334</td>
<td>2.126</td>
<td>.1254</td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups: RATERS</td>
<td>87</td>
<td>66.8333</td>
<td>.7682</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>70.1000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 12

ANOVA for KNOWING TO WHICH DIMENSION THE BEHAVIOR WAS RELATED

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups: CONDITION</td>
<td>2</td>
<td>1.0889</td>
<td>.5445</td>
<td>1.207</td>
<td>.3039</td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups: RATERS</td>
<td>87</td>
<td>39.2333</td>
<td>.4510</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>40.3222</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7 presents the mean responses, by condition, to the first five questions on the participant questionnaire. The left hand column presents an abbreviated statement of what the question asked. Table 8 presents the significance levels for t-tests comparing pairs of treatment means. Tables 9 through 13 present the analyses of variance which were conducted in order to test differences in responses to these items across conditions. While the ANOVAs and t-tests both compare the treatments by an analysis of their means, they present the data in slightly different forms. Whereas the ANOVAs provide a summary index of variability among several conditions, the t-tests provide comparisons of means for pairs of conditions.
Since the research hypotheses separate predictions for access to and use of the dimension-related behavior lists, the t-tests were included here to enable specific paired comparison of conditions. While both Wherry and this investigator see these as related conditions with increasing potency for improving the quality of ratings, strictly speaking, they are not steps in a single previously identified continuum. For these reasons, both types of analysis are included here.

None of the conditions (1, 2, 3) differ in the self-reported level of experience of raters. They all report limited experience with rating despite the supervisory job experience they report. (See Table 6.) (On the rating scales: 2=very little experience; 3=moderate.) Ss in conditions 2 and 3 were significantly more inclined to feel that the list of dimensions helped them than were raters in Condition 1. Marginally significant differences (i.e. alpha level greater than .05 but less than .10) were found between the responses of Condition 1 and Condition 2 Ss to questions concerning whether (1) they felt more prepared, (2) they knew what dimensions the behavior they saw was related to, or (3) the dimensions represented different characteristics. Raters in conditions 2 and 3 did not differ in their self perceptions on these questions. Raters in Condition 3
also felt the list helped, and felt more prepared than did raters in Condition 1.

A third aspect of the manipulation check involved the following section on the participant questionnaire.

IF YOU WANTED TO IMPROVE YOUR ABILITY TO ACCURATELY RATE THE PERSONS' PERFORMANCE WHICH WOULD YOU MOST LIKE TO HAVE? (Rank order items; start with 1 to indicate the most preferred.)

____ Better definitions of the characteristics
____ More characteristics to rate
____ Fewer characteristics to rate
____ More specific rating scales
____ More opportunity to observe the person(s) rated
____ More information about what behavior to pay attention to

Table 14 presents data on the responses of raters by condition. The data are presented in this raw form because the mean rankings and standard deviations of the means can provide diagnostic information for interpreting the effects of the different conditions upon the raters' perceptions.
# TABLE 14
Forced ranking of rater choices to improve their rating

<table>
<thead>
<tr>
<th>ITEM*</th>
<th>CONDITION 1</th>
<th></th>
<th>CONDITION 2</th>
<th></th>
<th>CONDITION 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean  s.d.</td>
<td>n</td>
<td>Mean  s.d.</td>
<td>n</td>
<td>Mean  s.d.</td>
<td>n</td>
</tr>
<tr>
<td>Better definitions</td>
<td>1.5  .8</td>
<td>29</td>
<td>3.0  1.2</td>
<td>26</td>
<td>2.3  1.5</td>
<td>29</td>
</tr>
<tr>
<td>More characteristics</td>
<td>4.7  1.5</td>
<td>28</td>
<td>5.2  1.3</td>
<td>25</td>
<td>5.1  1.6</td>
<td>25</td>
</tr>
<tr>
<td>Fewer characteristics</td>
<td>4.7  1.2</td>
<td>26</td>
<td>4.3  1.5</td>
<td>26</td>
<td>4.1  1.6</td>
<td>27</td>
</tr>
<tr>
<td>More specific scales</td>
<td>4.1  1.4</td>
<td>28</td>
<td>3.8  1.3</td>
<td>25</td>
<td>3.1  1.2</td>
<td>27</td>
</tr>
<tr>
<td>More opportunity to observe</td>
<td>2.8  1.5</td>
<td>29</td>
<td>1.6  1.1</td>
<td>26</td>
<td>2.5  1.4</td>
<td>28</td>
</tr>
<tr>
<td>More information about behavior</td>
<td>3.0  1.1</td>
<td>29</td>
<td>2.7  1.3</td>
<td>26</td>
<td>3.4  1.6</td>
<td>27</td>
</tr>
</tbody>
</table>

*Abbreviated form of item; see preceding copy of question for complete wording of items.
Raters in Condition 1 unequivocally would like better definitions. It tops their list and agreement among them is strong. The preferences and levels of agreement on them vary by condition.

The last piece of the manipulation check to be analyzed was the section which asked raters to answer the question: "What did you take into consideration when rating a person on: Initiative........... Command Influence............. Informed Judgment.." (see appendix for response format). These three constructs were chosen as a sample of the rating dimensions employed by all raters. It was assumed that if raters focused more on specific dimension related behaviors when they had had the opportunity to read or use behaviorally clarified dimension lists then that focus on behavior and specific activities of the ratee should be reflected in the raters reports of what they took into consideration. This is the supposition at the heart of supplementary hypothesis 1.

HS1: The criteria which raters report taking into consideration when rating an individual's performance will vary in content (e.g., specificity, behavioral language, distinctive character) depending upon the treatment group to which the raters belong.

In order to analyse the differences between responses from raters in the different conditions, a scoring scheme
had to be developed. To this end the investigator identified a set of categories which might conceivably reveal differences in the definitions and considerations reported by raters from different conditions. An a priori set of content coding rules were developed and adhered to in coding the raters' responses. As indicated previously, without a pre-test set of responses these analyses provide only tentative information. In consideration of that fact the investigator did not rigorously design a blind coding experience. (All response sheets were already labelled with condition specific identification numbers.) The investigator did, however, alternate between responses from different conditions and periodically calculated summary statistics to try to make sure that coder's judgments seemed to be made in a consistent manner without any secular trends. The content coding rules and examples of coded passages can be found in appendices H and I respectively. Table 15 presents a summary of the raw data from the content coding. Tables 16 through 20 present the analyses of variance which were conducted to compare the coding across conditions. One of the assumptions underlying the design of these experimental treatments was that the behavioral specificity of raters' schemas for classifying their observations into various dimension ratings would increase across conditions. The content coding categories were intended
to provide a set of measures of those differences. The number of action statements, references to specific actual behavior observed, and non-behavioral statements of the rating constructs in the definitions written by the Ss differed significantly by condition. Their definitions did not differ in length or in terms of references to the outcomes of the ratees' behaviors.
### TABLE 15

Content coding: The nature of the data raters report considering when making ratings

<table>
<thead>
<tr>
<th>CODING CATEGORY</th>
<th>Mean Number of Responses Fitting Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Condition 1</td>
</tr>
<tr>
<td>Words</td>
<td>57</td>
</tr>
<tr>
<td>Action statements</td>
<td>1.6</td>
</tr>
<tr>
<td>Specific actual behavior mentioned</td>
<td>.1</td>
</tr>
<tr>
<td>Reference to outcomes of behavior</td>
<td>.9</td>
</tr>
<tr>
<td>Non-behavioral statements of construct</td>
<td>2.8</td>
</tr>
</tbody>
</table>

*Note: number of complete responses within condition

### TABLE 16

ANOVA for CONTENT CODING OF DIMENSION DEFINITIONS: # of words

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>F Ratio</th>
<th>F Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups: CONDITION</td>
<td>2</td>
<td>3077.4</td>
<td>1538.7</td>
<td>2.186</td>
<td>.1191</td>
</tr>
<tr>
<td>Within groups: RATERS</td>
<td>79</td>
<td>55606.4</td>
<td>703.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>58683.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 17
ANOVA for CONTENT CODING OF DIMENSION DEFINITIONS:
# of action statements

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups:</td>
<td>2</td>
<td>58.5308</td>
<td>29.2654</td>
<td>9.594</td>
<td>.0002</td>
</tr>
<tr>
<td>CONDITION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups:</td>
<td>79</td>
<td>240.9812</td>
<td>3.0504</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>299.5119</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 18
ANOVA for CONTENT CODING OF DIMENSION DEFINITIONS:
Specific actual behavior referenced

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups:</td>
<td>2</td>
<td>18.7617</td>
<td>9.3808</td>
<td>3.390</td>
<td>.0387</td>
</tr>
<tr>
<td>CONDITION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within groups:</td>
<td>79</td>
<td>218.8747</td>
<td>2.7673</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>237.3778</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 19

ANOVA for CONTENT CODING OF DIMENSION DEFINITIONS:
Refers to outcomes

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONDITION</td>
<td>2</td>
<td>2.9770</td>
<td>1.4885</td>
<td>2.156</td>
<td>.1225</td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td>79</td>
<td>54.5351</td>
<td>.6903</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>57.5121</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 20

ANOVA for CONTENT CODING OF DIMENSION DEFINITIONS:
Non-behavioral statement of construct

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARES</th>
<th>F RATIO</th>
<th>F PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONDITION</td>
<td>2</td>
<td>53.9666</td>
<td>26.9833</td>
<td>10.372</td>
<td>.0001</td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>groups:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RATERS</td>
<td>81</td>
<td>210.7356</td>
<td>2.6017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>264.7021</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Hypothesis S2**

Participants in conditions 3 and 4 were instructed to use their lists of dimensions and associated behaviors as a checklist: "EACH TIME one of the two officers demonstrate a behavior related to one of the performance dimensions MAKE A CHECK MARK under his name next to the related behavior". Unfortunately, not all of the participants followed the instructions precisely. Only 16 out of 59 participants in the two conditions kept a raw frequency count of the behaviors demonstrated. The remaining Ss apparently checked a behavior if it was demonstrated by the ratee, but did not check if the behavior was demonstrated twice or more often. It was impossible to tell which Ss had seen a behavior only once and which had checked it only once to indicate that they had seen it at least one time. Consequently the checklist scores for conditions 3 and 4 were calculated to indicate how many of the behaviors within a dimension were observed (NOT how many times they were observed). This scoring method is consistent with the conceptual development of the dimensions and the logically related sets of behaviors which could be considered alternate manifestations of the construct/dimension. It is also consistent with the prescribed practice for scoring written material for
evidence of one of the three social motives—need for power, need for affiliation, need for achievement (see Atkinson, 1958); this type of competency/dimension scoring was derived as a variant of motive scoring practices. Most importantly, the scoring approach employed provides a consistent basis on which to compare Ss in terms of how many aspects of each dimension they picked up while observing the ratees.

Supplementary hypothesis S2 predicted that individuals who know that they will later have to make ratings will score observed behavior using checklists in the same way as those Ss who are not asked to make ratings. This hypothesis was tested by simply comparing the means of dimension scores made by raters in Condition 3 with those made by raters in Condition 4.

There were no significant differences between the mean checklist scores from Ss in Condition 3 and those from Ss in Condition 4. A table of the means and t-values are listed in Appendix J.

As noted earlier the checklist scoring process and scores it generates are not comparable to the rating process and dimension ratings. Nonetheless, in an effort to provide parallel analyses and comparisons of conditions similar to those conducted to compare
conditions 1, 2 and 3, correlational analyses were conducted on the checklist scores from Condition 3 and Condition 4. The analyses were identical to the earlier analyses except that the data points were scores not ratings. (See Appendix A for definitions that make distinctions between the two kinds of assessments.) The results of these analyses are presented in Tables 21 and 22.

**TABLE 21**

HS2: Scores from Ss in Condition 3 will not differ from scores from Ss in Condition 4—HALO

<table>
<thead>
<tr>
<th></th>
<th>CONDITION 3 (n=60)</th>
<th>CONDITION 4 (n=58)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.36</td>
<td>.32</td>
</tr>
</tbody>
</table>

Note: No significant differences were found in rs across ratees or condition

**TABLE 22**

HS2: Scores from Ss in Condition 3 will not differ from scores from Ss in Condition 4—INTERRATER RELIABILITY

<table>
<thead>
<tr>
<th></th>
<th>CONDITION 3 (n=30)</th>
<th>CONDITION 4 (n=29)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.51</td>
<td>.57</td>
</tr>
</tbody>
</table>

Note: No significant differences were found in rs across ratees or condition
The mean intercorrelations were compared using \( r \) to \( Z \) transformations and then a comparison of the resulting standardized \( z \) score to the normal probability curve. There were no significant differences in \( rs \) across conditions. Since the scoring and rating data are not comparable, no firm conclusions can be drawn concerning the relative reliability or halo in dimension assessments from one versus the other process.
DISCUSSION

This study was intended to test the propositions contained in Wherry's Theorems 6 and 8, i.e. that providing raters with access to behaviorally clarified rating dimensions, or requiring them to use lists of dimension-related behaviors as checklists would be reflected in more accurate ratings. The specific choices of treatments and dependent variables were derived after consideration of the various bodies of literature cited, and on the basis of current practice in performance rating within organizations. In 1980 Landy and Farr commented that "we must learn much more about the way in which potential raters observe, encode, store, retrieve and record performance information....we probably have gone as far as we can in improving rating formats" (p.69). This study held rating format constant across conditions and, instead, provided treatments which supplied a ten dimension framework for those ratings in 3 different ways. The behavioral specificity of those dimension lists and their use were predicted to result in both a decrease of halo and an increase in interrater reliability across conditions. The explicit hypotheses predicted that raters who had access to lists of behavior related to dimensions would produce better ratings judged
by these two criteria and that actual use of those lists as checklists would result in ratings that were better still.

As noted from the beginning of this document, many researchers in the fields of performance appraisal and psychological measurement have made a case for developing a specific shared understanding of rating dimensions among those who must use them to assess the observed performance of others. Finkle and Jones (1970) point out that development of a shared language allows judgements to be formulated, compared, and communicated accurately. On page 2 of this document the focus of this study is identified as the "clarification of dimensions as constructs representing aspects of performance that can be observed." That clarification was seen as integral to the quality of ratings since, as Wherry noted (1982), accurate ratings are a function of the ratee's performance, observation of that performance, and recall. The three formats of the 10 rating dimensions presented to raters were chosen to vary on the degree to which they might clarify the rating constructs for raters. Clarification was to be accomplished by presenting a specific set of statements indicating what kinds of behavior would be considered examples of the dimension in use on the job. Since much of the assessment that goes on in organizations, classrooms and clinics follows observation, and since it is often hard to get comparable measures of psychological
variables unless they have been defined in observable terms, **behavioral** clarification was selected as most appropriate.

Cohen and Ebbeson observed that the processing of a stream of behavior is influenced by the rater's prior knowledge and expectations about the interrelationships of behaviors, organized into a schema. Likewise, when Wherry first published his theory of rating in 1952 he noted that the rater's recall of information about observed behavior was based on the schema that was used to organize his/her memory. That schema may be considered a stereotype, a preconceived notion (DeNisi et al., 1984), or an implicit personality theory (Feldman, 1981). The raters in the three conditions in this study were presented with lists of dimensions that differed in behavioral specificity, and therefore, varied in the degree to which they were presumed to clarify and standardize the rating schema for use by raters. Condition 1 raters were provided with a list of dimensions that allowed for considerable variability in the way raters chose to interpret and use the rating constructs as they observed, categorized and evaluated ratee behaviors. Condition 2 raters were provided with a specific list of dimension-related behaviors as a standard framework which they could chose to use for observing, categorizing and evaluating the same
behavior. Condition 3 raters were further instructed in how to use the schema provided to observe, categorize, and store information on the behavior they observed, in order to later evaluate the ratees.

This study demonstrates that providing raters with access to schemas that varied in behavioral specificity, as operationalized in this study, did not have the predicted effects. Prescribed use of a behaviorally specific dimension list was, however, associated with lower halo. No significant differences between conditions were found for the dependent variable interrater reliability.

Despite the relative lack of predicted results, this study provides fertile ground for generating hypotheses about what factors might have positive impact on the information processing done by individuals who observe and rate the performance of others and, consequently, might be reflected in the halo and interrater reliability of ratings they make. Specifically, active use of a concrete, behavioral schema during observation is associated with lower halo. It is possible that both levels of halo and of interrater reliability could be improved by providing (1) increased time for raters to be exposed to the appropriate rating schema, and (2) training involving a
variety of learning experiences including the use of that schema. Specific observations on each of these points are included in the following discussion of the results from the analyses of each dependent variable and the analyses of the supplementary data which served as a manipulation check.

Dependent variable: Halo

ANOVA Analysis

Halo has been defined as "low intercategory variance" of the ratings made by an individual rater making multiple assessments of one ratee (Cooper, 1981). The standard deviation of those ratings, therefore, is one possible measure of halo. The analysis of variance in this study was performed using the standard deviation of each rater's ratings as the dependent variable. A similar analysis was performed by Borman and Dunnette (1975); they performed an analysis of variance using the standard deviation of several ratees' ratings on one dimension as a measure of the raters' ability to differentiate among the performances of various ratees. In this investigation the dependent variable was the standard deviation of the ratings made by one rater when rating a ratee on 10 dimensions. It was used as a measure of the rater's ability to differentiate among 10 different aspects (i.e., dimensions) of a ratee's performance.
Hypothesis 1 predicted that the halo in ratings made by the raters in Condition 1 who received only lists of the dimension names, would be greater than that in ratings from Condition 2 or 3 in which raters received the detailed behavior lists. Likewise, Hypothesis 3 postulated that halo would be significantly lower for ratings made by raters who used the behaviorally clarified dimension list as a checklist to record what they observed as they watched the ratees' job performance. Thus, taken together these hypotheses predicted a stepwise decrease in halo across conditions.

A one way ANOVA tested for treatment effects across conditions 1, 2, 3. The significant F for the ANOVA based on the standard deviations of ratings indicated that there was a significant main effect for conditions, i.e. halo varied by condition. The means for the dependent variable increased in the predicted direction. This ANOVA, therefore, provided support for the linear model implied when hypotheses 1 and 3 are considered together. Further analyses of the standard deviations of ratings by condition provided greater insight into the relative extent of halo in dimension ratings from the three conditions.

Hypothesis 1 addressed the issue of whether giving raters access to behaviorally clarified
dimension descriptions would be reflected positively in psychometric properties of the ratings. Wherry was suggesting such an effect in his Theorem 6 (1982). He proposed that "the rater will make more accurate ratings when (s)he has been forewarned concerning the types of activity to be rated since this will facilitate his (her) more properly focusing attention on such pertinent behavior" (p.533). Hypothesis 1 was separated from Hypothesis 3 in order to cleanly address this issue without reference to what form the access to dimensions takes (i.e. raters in both conditions 2 and 3 had access to the behavior lists although only Condition 3 raters used them as checklists.).

The data reported in Table 3 present the confidence intervals used to compare the level of halo (measured as standard deviation of ratings) in ratings made by raters who did not have access to such lists. Halo did not significantly decrease from Condition 1 to Condition 2, therefore the data did not support the investigator's contention that significant reductions of halo may be obtained by simply providing raters with access to lists of dimension-relevant behaviors.

One more common, albeit more complex, approach to reduction of halo in ratings has been to design and use BARS type rating instruments. The BARS approach not only
provides individual raters with access to lists of behaviors and the dimensions for which they can be considered evidence, but also requires the raters to use those behaviors as referents for their ratings. As indicated previously, there is mixed evidence that the BARS approach can be relied upon to consistently produce ratings with low levels of halo. The research literature has not convincingly demonstrated the merits of that approach over other, simpler approaches which also focus the rater's attention on dimension-related behavior (Bernardin 1977). The treatment selected for Condition 2 was intended to demonstrate that a simple approach could effect significant reductions in halo by focusing the raters on dimension-relevant behaviors. Prior to rating, raters in Condition 2 were provided with a list of the behaviors related to each of the 10 dimensions and were given a BRIEF period of time to review the list. The impact of the treatment on the rater's focus was apparently not sufficient to produce a significantly lower level of halo than that found in Condition 1. The experimental design appears to have overcorrected for the complexity of a sophisticated approach to scaling behaviors and consequently to have provided a treatment that was simple, but failed to significantly effect raters' ability to differentiate among the 10 dimensions when rating the
officers whose performance they observed.

In Theorem 6 of his rating theory Wherry postulated that "the rater will make more accurate ratings when he has been forewarned concerning the types of activity to be rated since this will facilitate his more properly focusing attention on such pertinent behavior" (1982, p.533). Tables 15 through 20 present data from the manipulation check. They show that the raters in conditions 2 and 3 who were "forewarned" about what behavior was relevant for their ratings, did in fact, describe a sample of dimensions in more behavioral terms, using action statements and specific references to what they observed, significantly more often than did the raters who had not received lists of dimension-related behaviors prior to observing the behavior to be rated. The treatment for Condition 2 accomplished the goal of focusing the raters' attention on dimension-related behavior. It did not result in more accurate ratings in terms of halo. Therefore, while Wherry's Theorem 6 may be accurate in postulating the conditions necessary to improve the level of halo in ratings, those conditions (as operationalized in this study's Condition 2) do not appear to be sufficient for halo reduction.

Perhaps it is the actual use of individual dimension-
related behaviors as referents while rating that produce the benefits common to successful applications of the BARS approach and of simpler rating formats such as summated scales. Raters in Condition 3 used their lists of dimension-related behaviors as checklists. On the basis of a comparison of the means for conditions 2 and 3 it appears that using the behaviorally specific dimension lists to keep track of behavior while observing a person's job performance has potential as a means of reducing rating halo. Condition 3 ratings had significantly less halo than did Condition 1 or Condition 2 ratings. Wherry predicted that if the observer were "furnished with an easily accessible checklist of objective cues for the evaluation of performance to which (s)he can frequently refer (s)he should be better able to focus his(her) attention properly" (1982; Theorem 8; p. 534).

Raters in all conditions were asked to respond to the question "Did the list help you to evaluate the officers' performance?" Raters in both conditions 2 and 3 felt that their lists of dimension-related behaviors were significantly more helpful than did raters in Condition 1 who had received lists of dimension names only. When later asked to indicate what they took into consideration when making ratings on a sample of three of the
dimensions, raters in conditions 2 and 3 gave more
behavioral and specific responses than did raters in
Condition 1. Thus, as indicated previously, raters in
conditions 2 and 3 did focus their attention on objective
and relevant cues for rating the officers on the 10
dimensions. Theorem 8, however, postulates that raters
who use that list as a checklist will "be better able to
focus their attention properly". "Better" than what?..or
whom?..measured how? If halo is the dependent variable on
which we expect to see improvements in rating quality,
and the standard deviation of ratings is the chosen
measure of halo, then raters in Condition 3 rate "better"
than do raters from Condition 2 who had the same lists of
behaviors but did not use them as checklists. Thus, the
results of the analysis of variance and analyses of the
manipulation check data in this study support both
Wherry's Theorem 8 and Hypothesis 2 of this investigation.

Access to behavioral examples of how dimensions
are demonstrated appears to have affected the raters'
perceptions of what was relevant for consideration when
rating the officers they had observed. What could have
increased the impact of those clarified perceptions on
the level of halo among ratings? Obviously, USE of the
schema provided (in the form of a checklist for observed
behavior) was associated with a significantly lower level
of halo. Use of the checklist may not provide optimal reduction of halo however. A series of studies by Bernardin and his colleagues explored the relative merits of diary keeping (i.e. raters keeping an ongoing record of observed behavior, similar in concept to a checklist) and of training of raters in order to reduce rating errors such as halo (Bernardin and Buckley, 1981; Bernardin, 1978; Bernardin and Walter, 1977). Bernardin and Buckley concluded that the best approach to improving rating quality included diary keeping, training raters to try to establish a common frame of reference on what constitutes effective performance on a rating dimension, and practice in using the rating procedure under difficult but realistic rating circumstances (1981). Other researchers have also explored the relative merits of different approaches to rater training for reduction of rating errors such as halo. Latham and Wexley (1981) concluded that the best rater training included practice in observing and evaluating behavior using the rating procedure.

The common element in both Bernardin and Buckley's (1981) and Latham and Wexley's (1981) prescriptions for improving ratings is the emphasis on actual rater use of the rating constructs during training. Wherry (1982) contends that any factor which will heighten a rater's sensory experience with the ratee and his/her performance
will increase the influence of actually observed behavior upon ratings made as against the influence of bias or error in the act of perceiving. That is the rationale behind including the use of checklists as a condition in this study. It is also one reason why experiential training in the use of rating procedures might be an advantageous addition to either condition 2 or 3. Since no training was included in this design, the actual benefits which training might add to these clarifications of dimensions is not known.

Correlational Analysis

The correlational analysis of halo revealed an interesting pattern but no significant differences across condition when comparing levels of halo measured this way. These analyses were conducted by correlating the ratings on each dimension (n=60 ratings; 2 for each of 30 raters per condition) with ratings on each of the other dimensions in turn. What this provided was a set of correlations between dimensions across raters. It was a gross measure of halo in that it did not sort out the interrelationships of dimensions within each set of 10 ratings for one ratee.

While results of the correlational analyses are disappointing in that there were no significant differences between conditions in support of either H1 or H3, this analysis provides additional useful information.
correlation matrices from which the mean r values in Table 3 were calculated are presented in Appendix G. Within the confines of this study, these data provided the only means for inferring the construct validity of the dimensions. The construct validity of the dimensions will not be addressed in detail in this discussion. Suffice to say that the patterns of intercorrelations were consistent with their definitions, with the realities of the jobs they were used to assess, and with their known relations to other variables. For example, Empathy negatively correlated with many of the other variables. It was selected as a dimension from a list of the competencies of a lower level management job; it rarely appears as a positive characteristic of performance in senior level jobs. Note also that the highest intercorrelations in the matrix of correlations from the total sample were among Planning, Monitoring for Results and Conscientious Use of Discipline. All of these dimensions are conceptually related to the well established construct nAch.

Dependent Variable: Interrater Reliability

Hypothesis 2 predicted that the interrater reliability of dimension ratings would be less for Condition 1 than for either Condition 2 or Condition 3. This prediction was tested using a treatment design under which only raters
in the latter two conditions had access to lists of rating dimensions clarified with descriptions of what ratee behaviors were relevant to each dimension. Additionally, Hypothesis 4 predicted that actual use of the behavior checklists to record observations would be associated with greater agreement among raters. Table 5 presents summary figures indicating the experimental level of interrater reliability by condition. Reliability figures appear to decrease from Condition 1 to 2, and then to increase from Condition 2 to 3. The reliability indices for conditions 1 and 3 are more nearly equal than are the reliability figures for other pairs of conditions. None of these differences, however, proved to be statistically significant.

While the differences by condition did not turn out to be significant, the results provide an interesting and unanticipated pattern. Moreover, that pattern parallels one reported by Borman and Dunnette (1975). In their study they compared 3 different rating systems. They were:

(1) trait ratings (similar to the rating task presented to Condition 1 here),

(2) unanchored ratings made by Ss who had received not only titles but also definitions of the rating dimensions (similar to condition 3), and
(3) ratings from anchored scales with behavioral anchors at scale points (Although different in format from Condition 3 here, this condition also provided an increased exposure to and use of behavioral referents).

Their trait ratings were made on a different set of dimensions than the other two sets of ratings. Their anchored scale condition was different from Condition 3 in this study. Nonetheless, the studies are quite similar. Borman and Dunnette found that interrater agreement decreased from trait to unanchored ratings and then increased for anchored scale ratings. Interrater reliabilities for trait and anchored ratings were nearly identical. They interpret this finding in terms of the familiarity of raters with the trait format (an existing set of traits assessed in the organization). Their findings suggest that the findings from this study are not idiosyncratic. Borman and Dunnette's results add weight to the contention that the pattern of nonsignificant results in this investigation deserves further study.

The most plausible explanation for the pattern of reliabilities reported in Table 5 is that they occurred by chance. Significance tests are, after all, used to eliminate chance differences from consideration as real effects. Because the pattern of results is not unique to this study, however, further research is warranted on these hypotheses and on other factors which may effect the reliability of ratings made under similar conditions.
The power of this particular experimental design was low; the first of any follow up studies should replicate the essential elements of this study with a larger experimental sample. The second such study should attempt to strengthen the treatments. Despite the lack of results found here the investigator is not yet convinced that the logic behind these hypotheses was essentially faulty. If neither of these follow up studies were to confirm the original hypotheses, then one could more firmly conclude that the reasoning behind them was faulty, that the predicted effects are too small to pass tests of significance, or that some other mechanism is operating which interacts with the clarification of rating dimensions for raters. All of these conclusions about the merits of these hypotheses could be studied further. Likewise, if the results from a parallel study with a larger sample were to find significant differences between conditions in a pattern that parallels that found both here and in Borman and Dunnette (1975), then further research would also be in order to explain that pattern. The following discussion suggests some directions for the research which could be developed to explore that pattern, or to better understand the lack of significant predicted results.

Failure to confirm the reliability hypotheses may have resulted from (1) underestimating the degree to which raters shared "preconceived notions" (DeNisi et al., 1984)
of how various behaviors relate to one another and (2) by overestimating the increased homogeneity of schemas induced by the treatments for conditions 2 and 3. Also since raters had never used the dimensions and dimension definitions in these ways, the interrater reliability figures may reflect a minimum level of reliability among all raters --- perhaps the raters are at the beginning of the learning curve when it comes to using and understanding that method of gathering information relevant to their ratings.

Although the raters from all conditions demonstrate similar levels of interrater agreement, the raters in each condition may be agreeing on the basis of different sets of implicit dimension definitions. The content coding data in Tables 13 -20 could be interpreted to support this notion. Raters in conditions 2 and 3 report making their judgments on the basis of less abstract, more observable, behavioral and specific aspects of ratee performance. If their self-reports are accurate then the ratings made by raters in conditons 2 and 3 should be more valid even though they are no more reliable than the ratings made by Condition 1 raters.

Raters in Condition 1 are as reliable as those in Condition 3. The halo of their ratings, however, significantly decreases across conditions. Recall that
the set of rating dimensions used in this study was chosen to include dimensions which should provide a range of ratings for each ratee. Note also that Appendix G shows high and low intercorrelations among dimension ratings, configured in a way that suggests that these dimensions are conceptually and operationally different from one another. Thus although no significant differences in reliability were found across conditions, the reliability of raters from conditions 2 and 3 would appear to be tied more closely to a reliance on actual observed behavior. It would be interesting to study the raters' operating schemas more closely in order to better understand why the ratings rooted in observed behavior are no more reliable than those based on the raters' non-behavioral schemas.

Another question that arises from viewing the pattern of reliability findings in Table 3 is: why does interrater reliability seem to fall off from Condition 1 to Condition 2? Again, this is not a significant difference in terms of statistical criteria. Perhaps that pattern arises from the nature of the learning situation for raters. As suggested before, all raters may be just starting up the learning curve, -- the task is new for all of them. This might be one avenue to explore in order to try to understand this pattern if it were to be replicated. Training and learning variables
may also provide fruitful explanations for a repeated lack of significant results.

Raters in Condition 2 were given a limited amount of time to read and review a set of dimension descriptions that were new to them. No discussion was allowed; no actual training took place. Perhaps this treatment allowed just enough time to disrupt their less behavioral, normative expectations about what the rating dimensions were intended to refer to, BUT not enough time to form new norms that could be easily considered in making judgments about behavior observed. The assumption underlying the experimental hypotheses was that the predicted increase in interrater reliability across conditions would occur as a consequence of raters learning and internalizing shared behavioral definitions of the rating constructs. Those shared definitions would operate as a schema that would sensitize raters to certain behaviors and facilitate categorization, storage and later recall of observations for the purpose of rating (DeNisi et al., 1984). Since the time allowed for raters to acquaint themselves with the rating constructs in this study was short, the likelihood of actually learning was not as great as had been implicitly hoped when the study was designed. The lack of time available to review the dimension lists may have contributed to the lack of significant differences between
all of the different conditions. Perhaps none of the raters had time to as fully acquaint themselves with the schemas as was necessary for those schemas to effect interrater agreement.

The research on learning styles provides one possible explanation for why a lack of appropriate learning opportunities may have contributed to the lack of significant findings for the interrater reliability hypotheses. According to theory and research in the area of experiential learning, most people need to be exposed to several different kinds of learning experiences in order to effectively learn (Kolb, 1976). The learning exercise for raters involved reading, but no observing of others using the materials, no practice in their use, and no opportunity to experiment with them. The time allotted for reviewing the lists of behaviors related to the rating dimensions did not provide a learning experience compatible with all the possible learning styles preferred by raters. It is an unconfirmed possibility that because none of the conditions provided training or sufficient variety of experiences to facilitate widespread learning of the new schemas, raters were unable to use them reliably when rating.

Without reference to Kolb, or the research on experiential learning, Latham and Wexley (1981) drew a similar conclusion about the essential elements required in any effort to reduce the errors raters make. They
concluded that only training programs that give ratees a chance to practice observing and rating behaviors,--with feedback on the accurate application of dimension ratings to those behaviors-- are likely to be effective in improving rating accuracy. In framing these hypotheses the investigator contended that the focus on accurately ascribing behavior to dimensions was the understated key to the success of those programs. It could be, however, that without the varied learning experiences included in that training, presenting lists and readings to raters in an attempt to clarify the relevance of particular behaviors to dimensions may not be sufficiently powerful to effect interrater reliability. Perhaps Wherry's theorems should be supplemented to indicate that the postulated advantages (1) of rating scale items that refer to easily observed behavior and (2) of the use of checklists of objective cues for evaluation of performance might be enhanced by means of multimethod training experiences. Research on the best means to provide learning for raters might enable researchers to better design treatments experimental treatments, and practitioners to design practical approaches, for introducing rating systems that are likely to produce reliable sets of assessments.
Manipulation check

The manipulation checks included in this investigation show that the treatments effected would affect the behavioral specificity, differentiation, and understanding of constructs by raters.

The groups of raters in the various experimental treatments had similar demographic profiles (see Table 6) therefore any apparent differences in results by condition could not be ascribed to sample differences on those demographic factors. The hypotheses in this investigation were based on the notion that those raters who were given behaviorally clarified descriptions of rating dimensions would produce psychometrically superior ratings because they would be better prepared to make accurate, consistent ratings thanks to the descriptive list which would help them to differentiate the dimensions, to understand what behaviors are relevant to judgments on which dimensions and to store the observations in memory for later recall and integration into ratings.

The manipulation check tried to tap these assumed correlates of the experimental treatments. (The results are presented in Tables 8-20.) In fact, raters in both conditions 2 and 3 felt the lists helped them and felt more prepared than their counterparts in Condition 1. Although the mean responses by Condition 2 and Condition 3 raters are essentially the same, a comparison by t-test
revealed only a marginally significant difference (alpha .10) between the mean responses from Condition 1 and Condition 2 on questions concerning the relation of behavior to dimensions or differentiation among dimensions. Nonetheless, when raters described what they considered when making ratings on 3 sample dimensions, raters from both conditions 2 and 3 responded by referring to more specific actions and behaviors than did Condition 1 raters (see Tables 15 through 20).

A more interesting observation involves the phenomenological perspectives of raters in Condition 2 and in Condition 3. No differences existed between the self-perceptions of raters in conditions 2 and 3 (see Tables 7 & 8). Nonetheless, the halo analyses show a pattern of Condition 3 raters outperforming Condition 2 raters. The comparison of means from the halo ANOVA revealed a significant difference between conditions 2 and 3. Also, condition 3 raters appeared more reliable although that apparent advantage over Condition 2 raters was not a significant one. The two groups of raters perceived that they were getting the same amount of help and insight from exposure to the clarified dimensions list, yet there was a tentative, but consistent, showing of greater impact on the quality of ratings from raters who actually USE those lists during observation.
The ranking data collected as part of the manipulation check provided another perspective on the operative assumptions of the hypotheses. When rank ordering options for changes in the rating process that might help them improve their ratings (see Table 14) raters in Condition 1 consistently chose "better definitions" as their preferred option. Raters from Condition 2 didn't place as much value on getting better definitions; perhaps this reflects satisfaction with the definitions they received. They would prefer "more opportunity to observe" the behavior they will be asked to rate. This observation is compatible with the investigator's premises that raters in Condition 2 and Condition 3 would focus more closely on dimension-related behaviors in order to make their ratings. Raters in Condition 3 ranked the options in a way that produced an approximate tie between preferences for "better definitions" and "more opportunity to observe". This suggests a dual recognition of (1) the need to observe behavior in order to rate accurately, and (2) the extent to which using a detailed checklist of dimension definitions helps the Condition 3 raters to feel that they have made accurate ratings.
Conclusion

In summary, the results of this investigation provided support for the predicted pattern of halo results. The analysis of variance did provide support for the hypotheses that halo would decrease across conditions. The reduction in halo was mainly attributable to use of a list of dimension-related behaviors as a checklist during observation of the person to be rated; access to such lists of behaviors did not have as much impact on halo of ratings. The correlational data did not confirm these hypotheses but were frankly less well suited to test the hypotheses since they aggregated the interrelationships among ratings across raters. The hypotheses which postulated that the treatments would have a similarly positive effect on interrater reliability across conditions were not supported. No significant differences on interrater reliability were found between conditions.

In conclusion, this investigation provided evidence for the conventional wisdom that focusing raters on dimension relevant behavior can help reduce halo in subsequent dimension ratings. Simply providing access to lists of dimension relevant behavior had only limited impact on halo (i.e., alpha level .10) use of that list while observing, however, was associated with significantly lower halo in the subsequent ratings.
The checklist method used in Condition 3 was a simple approach to rating that has seldom been cited in the recent research literature. A variant of this approach is used by some organizations. It is called competency coding, derives from motive scoring, and is not widely researched outside of that area of psychology. Nonetheless, at least in terms of its potential for reduction of halo, this approach deserves both further research and consideration when designing assessment processes for use in organizations.

Conclusions from the reliability analyses are much more speculative. There were no significant results from which to draw firm conclusions. The truest, simplest conclusion is that more research is needed. The interesting pattern of results is replicated elsewhere. Perhaps further research could be designed to explore the question of whether the pattern occurred by chance. If the pattern were replicated then research should proceed to explore the mechanisms by which it occurs.
The PROLOGUE provided an example of an assessment situation in which the assessors/raters:

(1) each had their own operational definitions of the assessment dimensions,

(2) tended to rate a person high or low across dimensions based either on demonstration of one important dimension or on other facts known about the ratee, and

(3) disagreed with each other's ratings of the job candidates.

Their ratings were fraught with the kinds of errors that the treatments in this study were intended to reduce.

These errors are more than abstract constructs. Their negative consequences can be visible and costly. Of the four candidates described in the prologue, Stan, Jean and Dennis were promoted. Tom was not. One year later only Jean was performing the new job well in terms of making changes and getting new systems on line and working.

Based on the results of this study and of the research literature in related areas the following recommendations could be made to the manager in charge of the next assessment center conducted in that organization.

*First, provide clear behavioral descriptions of the rating dimensions to each of your assessors. This should provide them with a common language for discussing ratees.

*Second, have the assessors use the list as a checklist as they observe the ratees. This should help reduce halo and may help increase intrarater reliability over that which is likely if you just give them the lists to read.

*Lastly, providing multi-method training experiences that help the raters to understand the list of dimensions may further enhance the positive consequences from their use.
REFERENCES


Bernardin, H.J. and Buckley, M.R. Strategies in rater training *Academy of Management Review,* 18=981, 6(2), 205-212.


REFERENCES (cont.)


REFERENCES (cont.)


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


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


Sackett, P. R. and Hakel, M. D. Temporal stability and individual differences in using assessment information to form overall ratings, Organizational Behavior and Human Performance, 1979, 23, 120-137.


Thurstone, L.L. Attitudes can be measured, American Journal of Sociology, 1928, 33, 524-529.
REFERENCES (cont.)


APPENDIX A: DEFINITIONS

The following definitions are adapted from Bales et al., 1979 (p. 241).

SCORING

—a method which tries to record behavior after the observer perceives it. Categories of behavior are set up in advance and the eventual scores for each category are frequency counts (i.e., sums) of related behaviors. The groups of behaviors are organized as indications of particular traits or characteristics of performers.

RATING

—a method of making evaluations after a period of interaction or observation of interaction. Ratings are also made for each of a predetermined set of categories (e.g., trait names; adjectives) on which the rater indicates the degree to which s/he observed that trait or characteristic of performance.
One hour and 10 minutes of *Twelve O'Clock High* was shown to the study participants. The film is actually one long flashback. The introductory scenes prior to the flashback were not shown. The rest of the film unfolds in chronological order.

When the film is started bombers are shown landing, having just returned from a bombing mission. The pilots, ground crew and equipment are all American. This is the 918th bomber squadron of the U.S. Army Air Corps, stationed in Archbury, England during World War II. Their mission is to disable German industry through daylight precision bombing of targets in occupied France and in Germany.

Colonel Keith Davenport (Gary Merrill) first appears at the airfield as the planes are landing. He greets one crew and offers them a ride back to the quonset hut where they will be debriefed on the mission just completed. He observes the debriefing and questions one crew after the emotional outburst of one of their members. Davenport asks that a recommendation for a medal of honor for that crew member be sent to headquarters.

When the squadron is alerted for the next day, he calls headquarters to question some specifics of the flight orders. When he gets no satisfaction Colonel Davenport goes to headquarters to appeal the orders. He argues with the general who issued the orders but is quickly told that the orders were made for a reason and that they will stand. Davenport complains that headquarters fails to recognize that the lost planes represent tired, overworked boys that have been stretched to the limit. The general proceeds to question Davenport about how many crews and planes he will have ready for tomorrow.

The general is Lt.Gen.Frank Savage (Gregory Peck). Although he firmly stands his ground he, nonetheless, appears to be a friend of Colonel Davenport's; he even gives him a gift of special flight boots. When Savage is later called to meet with his superior he is forced to speculate about why Davenport's group lost so many planes that day. He reluctantly concludes that Davenport is overidentifying with his men,—it's not bad luck, it's the squadron commander. Savage and his
superior go down to the squadron to get more information.

At the squadron, Davenport and his staff report on their activities that day. The navigator admits an error that caused them serious problems when they reached the target. Davenport takes responsibility for the error and then emotionally refuses to relieve the navigator as Savage recommends. The senior staff general relieves Davenport of command.

On the way back to headquarters the general requests that Lt. Gen. Savage take over command of the 918th; Savage accepts.

The next morning Savage arrives at the squadron. He quickly and emphatically enforces rules and regulations by chewing out subordinates and demoting people. He chastises the drunken adjutant. He demotes the air exec after severely reprimanding him for the role he played in Davenport's undoing and for dereliction of duty on several counts. He closes the officers' club "until further notice".

When he calls the entire squadron together, Savage immediately tells them that they will be doing practice missions every spare moment. He tells them that they had better not think they are anything special. If they intend to protect their own hides rather than put their all into the missions then they had better apply for transfer.

The medal of honor recipient arrives in Savage's office as a spokesman for the other pilots; they all want transfers. Later Savage strikes a deal with the adjutant to delay the paper work on those transfers. Practice missions begin and Savage rides along aboard the plane piloted by the air exec he demoted.

After one particular mission Savage calls another squadron meeting. He commends the group for several areas of improvement. He spends most of his time, however, individually chastising specific crew members for their personal failures during the mission. He concludes by asking if there are any questions. One officer rises to ask what has become of the transfer requests. Savage acts as if no one has spoken, puts on his hat, and leaves the room.
### APPENDIX C: RATING DIMENSION CHOICE

<table>
<thead>
<tr>
<th>Dimensions Considered</th>
<th>Dimensions Used in</th>
<th>PILOT</th>
<th>FINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Naval senior officer research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense of Responsibility</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>High Standards</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initiative</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Planning</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Effective Communication</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitoring for Results</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of Multiple Influence Strategies</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rational Persuasion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Command Influence</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Conscientious Use of Discipline</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Conceptualization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informed Judgment</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Positive Expectations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>From other competency studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathy</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Interpersonal Diagnosis</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Building Morale</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D:
Pilot Rating Data

<table>
<thead>
<tr>
<th>Dimension</th>
<th>DAVENPORT Condition*</th>
<th>SAVAGE Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td>2.5</td>
<td>1.7</td>
</tr>
<tr>
<td>INITIATIVE</td>
<td>31.</td>
<td>3.2</td>
</tr>
<tr>
<td>PLANNING</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>EFFECTIVE COMMUNICATION</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>MONITORING FOR RESULTS</td>
<td>3.5</td>
<td>3.4</td>
</tr>
<tr>
<td>USE OF MULTIPLE INFLUENCE STRATEGIES</td>
<td>3.7</td>
<td>3.4</td>
</tr>
<tr>
<td>COMMAND INFLUENCE</td>
<td>2.5</td>
<td>2.6</td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF DISCIPLINE</td>
<td>3.6</td>
<td>4.5</td>
</tr>
<tr>
<td>CONCEPTUALIZATION</td>
<td>3.7</td>
<td>3.4</td>
</tr>
<tr>
<td>INFORMED JUDGEMENT</td>
<td>3.0</td>
<td>2.6</td>
</tr>
<tr>
<td>POSITIVE EXPECTATIONS</td>
<td>3.1</td>
<td>3.3</td>
</tr>
<tr>
<td>EMPATHY</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td>2.2</td>
<td>2.3</td>
</tr>
<tr>
<td>BUILDING MORALE</td>
<td>2.8</td>
<td>2.2</td>
</tr>
</tbody>
</table>

*n=10 in each condition

**Ratings were made using a 1 to 5 scale on which 1 indicated "a great deal shown" and 5 indicated "little or none shown".
APPENDIX E: MATERIALS

ALL MATERIALS IN ORDER PRESENTED

ORAL INSTRUCTIONS

(These are the oral instructions to the participants as they enter the room and take their seats prior to opening their envelopes of materials.)

DISTRIBUTE MATERIALS, ALL MARKED WITH CODE NUMBERS, IN ENVELOPES,—HAND THEM OUT AT THE DOOR.

Please sit in the section which corresponds to the number on your envelope. DO NOT open your materials until asked to do so.

(Repeat as needed as people enter.)

ROOM WILL BE SPLIT IN QUARTERS AND MARKED OFF.

(When all are seated...) Good afternoon. McBer and Company is working with OIS and the Navy to study the ways in which officers evaluate other people's performance. This is something which you will have to do after you leave OIS and take on the responsibilities of officers.

The task you are about to undertake will provide information to the Navy and OIS which will help them to understand how officers make evaluations and may help them to train new officers.

Please open your envelope. Remove only the WHITE pages which are clipped to THE YELLOW SHEET labelled INSTRUCTIONS.

PLEASE READ AND FOLLOW THE INSTRUCTIONS ON YOUR MATERIALS.
INSTRUCTIONS GIVEN TO GROUPS 1 & 2

INSTRUCTIONS

NOTE: You have been divided into four (4) groups which received slightly different materials. Read ONLY your own materials. DO NOT TALK WITH THE OTHERS. You will all receive the same amount of time to review the instructions and materials you receive.

Officers are also managers who must observe and evaluate their subordinates' performance on the job. There are many different ways in which one can perform well or poorly on the job.

You will be viewing a film in which two (2) different persons perform various duties as senior officers with an air command. You will view a portion of the film and you will evaluate the performance of these two (2) people on ten (10) different dimensions of performance.

On the following page is a list of the ten (10) dimensions of performance. READ over this material and THINK ABOUT each of these different aspects of job performance. You will have fifteen (15) minutes to look over your materials.

When the fifteen (15) minutes are up the film will start. You will watch the film and OBSERVE the behavior of two (2) officers. After the film is stopped you will be asked to EVALUATE how well those two (2) officers performed on each of the job performance dimensions.

The film will begin in fifteen (15) minutes. It will run for one and one-half hours.
INSTRUCTIONS GIVEN TO GROUP 3

INSTRUCTIONS

NOTE: You have been divided into four (4) groups which received slightly different materials. Read ONLY your own materials. DO NOT TALK WITH THE OTHERS. You will all receive the same amount of time to review the instructions and materials you receive.

Officers are also managers who must observe and evaluate their subordinates' performance on the job. There are many different ways in which one can perform well or poorly on the job.

You will be viewing a film in which two (2) different persons perform various duties as senior officers with an air command. You will view a portion of the film and you will evaluate the performance of these two (2) people on ten (10) different dimensions of performance.

On the following page is a list of the ten (10) dimensions of performance. READ over this material and THINK ABOUT each of these different aspects of job performance. You will have fifteen (15) minutes to look over your materials.

When the fifteen (15) minutes are up the film will start. You will watch the film and OBSERVE the behavior of two (2) officers. You will be asked to keep track of how those two (2) officers perform each of the job performance dimensions. After the film is stopped you will be asked to EVALUATE how well those two (2) officers performed on each of the job performance dimensions.

The film will begin in fifteen (15) minutes. It will run for one and one-half hours.
INSTRUCTIONS GIVEN TO GROUP 4

INSTRUCTIONS

NOTE: You have been divided into four (4) groups which received slightly different materials. Read ONLY your own materials. DO NOT TALK WITH THE OTHERS. You will all receive the same amount of time to review the instructions and materials you receive.

Officers are also managers who must observe and evaluate their subordinates' performance on the job. There are many different ways in which one can perform well or poorly on the job.

You will be viewing a film in which two (2) different persons perform various duties as senior officers with an air command. You will view a portion of the film and you will evaluate the performance of these two (2) people on ten (10) different dimensions of performance.

On the following page is a list of the ten (10) dimensions of performance. READ over this material and THINK ABOUT each of these different aspects of job performance. You will have fifteen (15) minutes to look over your materials.

When the fifteen (15) minutes are up the film will start. You will watch the film and OBSERVE the behavior of two (2) officers. You will be asked to keep track of how those two (2) officers perform each of the job performance dimensions.

The film will begin in fifteen (15) minutes. It will run for one and one-half hours.
DIMENSION LIST GIVEN TO GROUP 1

PERFORMANCE DIMENSIONS

Read over this list of the dimensions of job performance. Take time to READ the list carefully and THINK ABOUT what each of the dimensions means.

You will observe the behavior of two (2) officers as they perform various duties during the film. Afterwards you will evaluate the officer's performance in terms of these ten (10) different aspects of job performance.

Other groups have different materials and instructions. You all have fifteen (15) minutes to review your own materials. NO TALKING, PLEASE; raise your hand if you have a question.

The film will begin shortly. Pay special attention to the behavior of COLONEL KEITH DAVENPORT (Gary Merrill) and GENERAL FRANK SAVAGE (Gregory Peck).

INFORMED JUDGMENT

COMMAND INFLUENCE

BUILDING MORALE

MONITORING FOR RESULTS

EMPATHY

CONSCIENTIOUS USE OF DISCIPLINE

PLANNING

INTERPERSONAL DIAGNOSIS

SENSE OF RESPONSIBILITY

INITIATIVE
DIMENSION LIST GIVEN TO GROUP 2

PERFORMANCE DIMENSIONS

Read over this list of behaviors which are related to the dimensions of job performance. Take time to READ the list carefully and THINK ABOUT what each of the dimensions means.

You will observe the behavior of two (2) officers as they perform various duties during the film. Afterwards you will evaluate the officer's performance in terms of these ten (10) different aspects of job performance.

Other groups have different materials and instructions. You all have fifteen (15) minutes to review your own materials. NO TALKING, PLEASE; raise your hand if you have a question.

The film will begin shortly. Pay special attention to the behavior of COLONEL KEITH DAVENPORT (Gary Merrill) and GENERAL FRANK SAVAGE (Gregory Peck).

INFORMED JUDGMENT

A person demonstrates INFORMED JUDGEMENT when s/he:

a. personally gathers information and identifies available facts in order to form an opinion or make a decision,
b. asks others to gather or provide information needed to form an opinion or make a decision,
c. uses data and information from experience to make a decision or draw a conclusion.

COMMAND INFLUENCE

A person demonstrates COMMAND INFLUENCE when s/he:

a. visits work areas, or otherwise makes self available or visible with the express purpose of showing interest, concern, or appreciation,
b. publicizes pleasure or displeasure with the group's or an individual's performance by tone of voice, demonstrative gestures, etc.,
c. creates or uses symbols to increase morale, loyalty, or sense of belonging (e.g., mottos, flags)
d. communicates standards and expectations by publicizing their enforcement.
BUILDING MORALE
A person demonstrates BUILDING MORALE when s/he:

a. encourages others in the face of setbacks,
b. expresses the need for keeping morale high,
c. gives recognition to others when they succeed.

MONITORING FOR RESULTS
A person demonstrates MONITORING FOR RESULTS when s/he:

a. actively observes work progress; seeks and collects performance information,
b. reviews finished products or final results for quality or accuracy,
c. evaluates training plans and activities, emphasizing proficiency and thoroughness.

EMPATHY
A person demonstrates EMPATHY when s/he:

a. describes other’s feelings, motives or situations and attempts to understand them,
b. acknowledges influences in a situation that may contribute to a person’s feelings.

CONSCIENTIOUS USE OF DISCIPLINE
A person demonstrates CONSCIENTIOUS USE OF DISCIPLINE when s/he:

a. uses a threat or a dramatic display of anger to coerce,
b. despite a possible concern for the individual’s future, has no reservation about exercising power in serious discipline cases or when harm to squadron, crew, or equipment appears likely,
c. enforces disciplinary standards.
PLANNING

A person demonstrates PLANNING when s/he:

a. plans beyond the demands of an immediate situation or problem,
   b. develops a specific plan of activities and milestones to reach a
gen,  
c. sets priorities on level of effort to be expended on various
activities,  
d. identifies obstacles to goal achievement,  
e. matches people to jobs to get the best performance,  
f. makes use of outside resources (programs, people, or funds) to
achieve goals.

INTERPERSONAL DIAGNOSIS

A person demonstrates INTERPERSONAL DIAGNOSIS when s/he:

a. describes the perspective of another person,  
b. anticipates people's reactions,  
c. interprets unstated meaning in an interpersonal situation,  
d. notices subtle emotional responses of others.

SENSE OF RESPONSIBILITY

A person demonstrates SENSE OF RESPONSIBILITY when s/he:

a. takes responsibility for safety and well-being of the squadron,  
b. takes actions to promote the well-being of families of squadron
members,  
c. takes responsibility for own or squadron's failures or problems,  
d. takes responsibility for unit's reputation or image.

INITIATIVE

A person demonstrates INITIATIVE when s/he:

a. introduces new ideas or procedures to the squadron,  
b. takes repeated action to overcome an obstacle or circumvent a problem,  
c. proposes, to people outside own unit, new ideas or better ways to
proceed,  
d. acts quickly or immediately to resolve problems.
CHECKLIST GIVEN TO GROUP 3; HALF WERE PRODUCED WITH NAMES IN REVERSED ORDER

PERFORMANCE DIMENSIONS CHECKLIST

Read over this list of behaviors which are related to the dimensions of job performance. Take time to READ the list carefully and THINK ABOUT what each of the dimensions means.

You will observe the behavior of two (2) officers as they perform various duties during the film. Afterwards you will evaluate the officer's performance in terms of these ten (10) different aspects of job performance.

While the film is running you should use this as a CHECKLIST to make a record of the officer's performance. EACH TIME one of the two officers demonstrates a behavior related to one of the performance dimensions, MAKE A CHECK MARK under his name, next to the related behavior or dimension. Keep track of DAVENPORT's and SAVAGE's behavior in separate columns since you will evaluate each of them separately after the film.

Other groups have different materials and instructions. You all have fifteen (15) minutes to review your own materials. NO TALKING, PLEASE; raise your hand if you have a question.

The film will begin shortly. Pay special attention to the behavior of COLONEL KEITH DAVENPORT (Gary Merrill) and GENERAL FRANK SAVAGE (Gregory Peck).

DAVENPORT

INFORMED JUDGMENT

_____ a. Personally gathers information and identifies available facts in order to form an opinion or make a decision

_____ b. Asks others to gather or provide information needed to form an opinion or make a decision

_____ c. Uses data and information from experience to make a decision or draw a conclusion

SAVAGE
COMMAND INFLUENCE

a. Visits work areas, or otherwise makes self available or visible with the express purpose of showing interest, concern, or appreciation

b. Publicizes pleasure or displeasure with the group's or an individual's performance by tone of voice, demonstrative gestures, etc.

c. Creates or uses symbols to increase morale, loyalty, or sense of belonging (e.g., mottos, flags, nautical language)

d. Communicates standards and expectations by publicizing their enforcement

- b. Expresses the need for keeping morale high

- c. Gives recognition to others when they succeed

BUILDING MORALE

a. Encourages others in the face of setbacks

b. Expresses the need for keeping morale high

c. Gives recognition to others when they succeed

MONITORING FOR RESULTS

a. Actively observes work progress; seeks and collects performance information

b. Reviews finished products or final results for quality or accuracy

c. Evaluates training plans and activities, emphasizing proficiency and thoroughness
DAVENPORT

EMPATHY

a. Describes other's feelings, motives or situations and attempts to understand them

b. Acknowledges influences in a situation that may contribute to a person's feelings.

CONSCIENTIOUS USE OF DISCIPLINE

a. Uses a threat or a dramatic display of anger to coerce

b. Despite a possible concern for the individual's future, has no reservation about exercising power in serious discipline cases or when harm to squadron, crew, or equipment appears likely

c. Enforces disciplinary standards

PLANNING

a. Plans beyond the demands of an immediate situation or problem

b. Develops a specific plan of activities and milestones to reach a goal

c. Sets priorities on level of effort to be expended on various activities

d. Identifies obstacles to goal achievement

e. Matches people to jobs to get the best performance

f. Makes use of outside resources (programs, people, or funds) to achieve goals
INTERPERSONAL DIAGNOSIS

a. Describes the perspective of another person
b. Anticipates people's reactions
c. Interprets unstated meaning in an interpersonal situation
d. Notices subtle emotional responses of others

SENSE OF RESPONSIBILITY

a. Takes responsibility for safety and well-being of the squadron
b. Takes actions to promote the well-being of families of squadron members
c. Takes responsibility for own or squadron's failures or problems
d. Takes responsibility for unit's reputation or image

INITIATIVE

a. Introduces new ideas or procedures to the squadron
b. Takes repeated action to overcome an obstacle or circumvent a problem
c. Proposes, to people outside own unit, new ideas or better ways to proceed
d. Acts quickly or immediately to resolve problems
PERFORMANCE DIMENSIONS CHECKLIST

Read over this list of behaviors which are related to the dimensions of job performance. Take time to READ the list carefully and THINK ABOUT what each of the dimensions means.

You will observe the behavior of two (2) officers as they perform various duties during the film. While the film is running you should use this as a CHECKLIST to make a record of the officer's performance. EACH TIME one of the two officers demonstrates a behavior related to one of the performance dimensions, MAKE A CHECK MARK under his name, next to the related behavior or dimension. Keep track of DAVENPORT's and SAVAGE's behavior in separate columns.

Other groups have different materials and instructions. You all have fifteen (15) minutes to review your own materials. NO TALKING, PLEASE; raise your hand if you have a question.

The film will begin shortly. Pay special attention to the behavior of COLONEL KEITH DAVENPORT (Gary Merrill) and GENERAL FRANK SAVAGE (Gregory Peck).

DAVENPORT

INFORMED JUDGMENT

a. Personally gathers information and identifies available facts in order to form an opinion or make a decision

b. Asks others to gather or provide information needed to form an opinion or make a decision

c. Uses data and information from experience to make a decision or draw a conclusion

SAVAGE
READ ALOUD TO ENTIRE GROUP

INTRODUCTION TO THE FILM

(This comes after the Ss spend 15 minutes reading over materials and just prior to starting the film.)

You will be viewing a portion of the film 12 O'CLOCK HIGH.

It is the story of the 918th squadron of the U.S. Air Command which served in England during World War II. The film was made in 1946. It presents many of the actual experiences of that command.

Focus on the following 2 officers as you view the film.

*COLONEL KEITH DAVENPORT (played by Gary Merrill)
  Davenport is the squadron commander as the film opens.

*GENERAL FRANK SAVAGE (played by Gregory Peck)
  He first appears when Davenport visits him in his office at headquarters (called Pinetree).

These two(2) men will be pointed out to you when they first appear.

As the film begins, the squadron is returning to its base after a bombing mission.
GROUP 4 INSTRUCTIONS WHEN OTHER PARTICIPANTS WERE RATING

INSTRUCTIONS

Officers in the other three groups have a specific rating task to complete at this time. You were asked to do something different.

While they complete their rating forms you can relax for a few minutes.

NO TALKING PLEASE!!!

In about 5 minutes you will be asked to complete one more form.

PLEASE wait to begin when everyone is told to do so.
Performance Evaluation: Davenport

Instructions

Think about what you observed as you watched Davenport perform his duties. Evaluate how he performed in terms of the each of the dimensions of performance listed below.

Circle the number which corresponds to your evaluation of his performance.

Be sure to circle one number for each performance dimension. Leave no rating scales blank.

<table>
<thead>
<tr>
<th>Informed Judgment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>quite a lot shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a moderate amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small amount shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very little or none shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Command Influence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>quite a lot shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a moderate amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small amount shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very little or none shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Building Morale</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>quite a lot shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a moderate amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small amount shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very little or none shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Monitoring for Results</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>quite a lot shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a moderate amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small amount shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very little or none shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Empathy</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>quite a lot shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a moderate amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small amount shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very little or none shown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Conscientious Use of Discipline

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>a great deal shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

## Planning

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>a great deal shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

## Interpersonal Diagnosis

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>a great deal shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

## Sense of Responsibility

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>a great deal shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

## Initiative

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>a great deal shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>a moderate amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>
PERFORMANCE EVALUATION: SAVAGE

Instructions

THINK ABOUT WHAT YOU OBSERVED AS YOU WATCHED SAVAGE PERFORM HIS DUTIES. EVALUATE HOW HE PERFORMED IN TERMS OF EACH OF THE DIMENSIONS OF PERFORMANCE LISTED BELOW.

CIRCLE THE NUMBER WHICH CORRESPONDS TO YOUR EVALUATION OF HIS PERFORMANCE.

BE SURE TO CIRCLE ONE NUMBER FOR EACH PERFORMANCE DIMENSION. LEAVE NO RATING SCALES BLANK.

INFORMED JUDGEMENT

1............2............3............4............5
a great deal shown quite a lot shown a moderate amount amount small shown shown little or shown shown very little or none shown

COMMAND INFLUENCE

1............2............3............4............5
a great deal shown quite a lot shown a moderate amount amount small shown shown little or shown shown very little or none shown

BUILDING MORALE

1............2............3............4............5
a great deal shown quite a lot shown a moderate amount amount small shown shown little or shown shown very little or none shown

MONITORING FOR RESULTS

1............2............3............4............5
a great deal shown quite a lot shown a moderate amount amount small shown shown little or shown shown very little or none shown

EMPATHY

1............2............3............4............5
a great deal shown quite a lot shown a moderate amount amount small shown shown little or shown shown very little or none shown
### CONSCIENTIOUS USE OF DISCIPLINE

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>quite a lot shown</td>
<td>a moderate amount shown</td>
<td>small amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

### PLANNING

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>quite a lot shown</td>
<td>a moderate amount shown</td>
<td>small amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

### INTERPERSONAL DIAGNOSIS

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>quite a lot shown</td>
<td>a moderate amount shown</td>
<td>small amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

### SENSE OF RESPONSIBILITY

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>quite a lot shown</td>
<td>a moderate amount shown</td>
<td>small amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>

### INITIATIVE

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>a great deal shown</td>
<td>quite a lot shown</td>
<td>a moderate amount shown</td>
<td>small amount shown</td>
<td>very little or none shown</td>
</tr>
</tbody>
</table>
PARTICIPANT QUESTIONNAIRE

Instructions

EVERYONE WANTS ACCURATE PERFORMANCE EVALUATIONS, THE NAVY DOES AND SO DO BOTH THE PERSON BEING EVALUATED AND THE PERSON DOING THE EVALUATING.

THE FOLLOWING INFORMATION IS NEEDED TO EXPLAIN THE DIFFERENCES AMONG OFFICERS DOING EVALUATIONS.

BE SURE TO ANSWER ALL OF THE QUESTIONS.

HOW MUCH EXPERIENCE HAVE YOU HAD IN RATING OR EVALUATING THE PERFORMANCE OF OTHERS?

1. no experience
2. very little experience
3. moderate experience
4. quite a lot of experience
5. a great deal of experience

YOU WERE GIVEN MATERIAL TO READ AND WERE THEN ASKED TO EVALUATE THE PERFORMANCE OF TWO (2) PEOPLE.

DID THE LIST HELP YOU TO EVALUATE THE OFFICERS' PERFORMANCE?

1. not at all helpful
2. a little help
3. some help
4. helpful
5. very helpful

HOW WELL PREPARED DID YOU FEEL?

1. not at all prepared
2. somewhat prepared
3. moderately prepared
4. well prepared
5. very well prepared

WHEN THE PERSONS BEING EVALUATED SAID OR DID SOMETHING DID YOU KNOW WHAT DIMENSION OF JOB PERFORMANCE THAT BEHAVIOR WAS RELATED TO?

1. had no idea what the behavior related to
2. was unsure what the behavior related to
3. had some idea what the behavior related to
4. was fairly confident what the behavior related to
5. knew exactly what the behavior related to

YOU HAD TO MAKE TEN (10) RATINGS FOR EACH PERSON. THE DIMENSIONS OF JOB PERFORMANCE YOU EVALUATED THEM ON SEEMED TO BE:

1. very different
2. somewhat different
3. somewhat similar
4. quite similar
5. the same thing with different names from one another

very somewhat somewhat quite the same
different different similar similar thing with
different names
YOU WERE ASKED TO EVALUATE THE PERSON’S PERFORMANCE IN TERMS OF SEVERAL DIMENSIONS OF JOB PERFORMANCE. HOW SURE WERE YOU THAT YOU MADE ACCURATE EVALUATIONS?

<table>
<thead>
<tr>
<th>OVERALL</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not at all sure</td>
<td>somewhat sure</td>
<td>moderately sure</td>
<td>quite sure</td>
<td>absolutely sure</td>
</tr>
</tbody>
</table>

HOW SURE WERE YOU THAT YOU HAD MADE ACCURATE RATINGS FOR EACH PERFORMANCE DIMENSION?

<table>
<thead>
<tr>
<th>INFORMED JUDGMENT</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>not at all sure</td>
<td>somewhat sure</td>
<td>moderately sure</td>
<td>quite sure</td>
<td>absolutely sure</td>
</tr>
<tr>
<td>COMMAND INFLUENCE</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>BUILDING MORALE</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>MONITORING FOR RESULTS</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>EMPATHY</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF DISCIPLINE</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>PLANNING</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>INITIATIVE</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
WHAT DID YOU TAKE INTO CONSIDERATION WHEN RATING A PERSON ON:

INITIATIVE ________________________________________________

__________________________________________________________

__________________________________________________________

__________________________________________________________

__________________________________________________________

COMMAND INFLUENCE _________________________________________

__________________________________________________________

__________________________________________________________

__________________________________________________________

__________________________________________________________

INFORMED JUDGEMENT _________________________________________

__________________________________________________________

__________________________________________________________

__________________________________________________________

__________________________________________________________
IF YOU WANTED TO IMPROVE YOUR ABILITY TO ACCURATELY RATE THE PERSONS' PERFORMANCE WHICH WOULD YOU MOST LIKE TO HAVE? (Rank order items; start with 1 to indicate the most preferred.)

______ Better definitions of the characteristics
______ More characteristics to rate
______ Fewer characteristics to rate
______ More specific rating scales
______ More opportunity to observe the person(s) rated
______ More information about what behavior to pay attention to

TO HELP US UNDERSTAND THE DIFFERENCES AMONG OFFICERS DOING EVALUATIONS. PLEASE ANSWER THE FOLLOWING QUESTIONS ABOUT YOURSELF.

CORPS (WRITE OUT NAME) _______________________________________
SEX: ________ male
____________ female

PRIOR SERVICE: ______ no; explain ______________________________
____________ yes

Have you taken LMET course before? ________ no
____________ yes

If yes, when and where? ________________________________________

Have you seen the film 12 O'Clock High before? ________ no
____________ yes

If yes, how many times? ____________

....... how recently? __________________

Have you held a supervisory job? ______ no
____________ yes
### APPENDIX F

**MEAN DIMENSION SCORES: BY RATEE, ACROSS CONDITION**

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>DAVENPORT N</th>
<th>MEAN</th>
<th>SAVAGE N</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFORMED JUDGMENT</td>
<td>90</td>
<td>3.0</td>
<td>90</td>
<td>1.7</td>
</tr>
<tr>
<td>COMMAND INFLUENCE</td>
<td>90</td>
<td>2.6</td>
<td>90</td>
<td>1.7</td>
</tr>
<tr>
<td>BUILDING MORALE</td>
<td>90</td>
<td>2.5</td>
<td>90</td>
<td>3.1</td>
</tr>
<tr>
<td>MONITORING FOR RESULTS</td>
<td>89</td>
<td>3.7</td>
<td>90</td>
<td>1.8</td>
</tr>
<tr>
<td>EMPATHY</td>
<td>90</td>
<td>1.3</td>
<td>90</td>
<td>3.7</td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF DISCIPLINE</td>
<td>90</td>
<td>4.4</td>
<td>90</td>
<td>3.7</td>
</tr>
<tr>
<td>PLANNING</td>
<td>90</td>
<td>3.8</td>
<td>90</td>
<td>1.5</td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td>89</td>
<td>2.5</td>
<td>90</td>
<td>2.6</td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td>90</td>
<td>1.9</td>
<td>90</td>
<td>1.9</td>
</tr>
<tr>
<td>INITIATIVE</td>
<td>90</td>
<td>3.4</td>
<td>90</td>
<td>1.4</td>
</tr>
</tbody>
</table>
### APPENDIX G:
INTERCORRELATION OF DIMENSION RATINGS

**TABLE 17:**
Total sample (n=130 sets of 10 ratings)

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>IJ</th>
<th>CI</th>
<th>BM</th>
<th>MR</th>
<th>E</th>
<th>CUD</th>
<th>P</th>
<th>ID</th>
<th>SR</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFORMED JUDGMENT</td>
<td>.55</td>
<td>.13</td>
<td>.62</td>
<td>-.33</td>
<td>.55</td>
<td>.64</td>
<td>.23</td>
<td>.28</td>
<td>.61</td>
<td></td>
</tr>
<tr>
<td>COMMAND INFLUENCE</td>
<td>.24</td>
<td>.43</td>
<td>-.20</td>
<td>.37</td>
<td>.43</td>
<td>.19</td>
<td>.16</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUILDING MORALE</td>
<td>ns</td>
<td>.38</td>
<td>-.13</td>
<td>ns</td>
<td>.39</td>
<td>.38</td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MONITORING FOR RESULTS</td>
<td>-.49</td>
<td>.70</td>
<td>.75</td>
<td>ns</td>
<td>.16</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPATHY</td>
<td>-.66</td>
<td>-.61</td>
<td>.26</td>
<td>.13</td>
<td>.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF DISCIPLINE</td>
<td>.82</td>
<td>ns</td>
<td>ns</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLANNING</td>
<td>.13</td>
<td>.13</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td>.20</td>
<td>NS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INITIATIVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ns means non-significant; r is greater than .05.
### Table 18:
Intercorrelation of ratings from Condition 1 (n=60 sets of ratings)

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>IJ</th>
<th>CI</th>
<th>BM</th>
<th>MR</th>
<th>E</th>
<th>CUD</th>
<th>P</th>
<th>ID</th>
<th>SR</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFORMED JUDGMENT</td>
<td>.71</td>
<td>ns</td>
<td>.78</td>
<td>-.55</td>
<td>.69</td>
<td>.74</td>
<td>.39</td>
<td>.47</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>COMMAND</td>
<td>ns</td>
<td>.48</td>
<td>-.42</td>
<td>.52</td>
<td>.55</td>
<td>.22</td>
<td>.24</td>
<td>.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFLUENCE</td>
<td>ns</td>
<td>.32</td>
<td>ns</td>
<td>ns</td>
<td>.58</td>
<td>.38</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUILDING MORALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MONITORING FOR RESULTS</td>
<td>-.53</td>
<td>.77</td>
<td>.84</td>
<td>.29</td>
<td>.30</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPATHY</td>
<td>-.65</td>
<td>-.69</td>
<td>ns</td>
<td>ns</td>
<td>-.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF DISCIPLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.81</td>
<td>.38</td>
<td>.38</td>
<td>.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLANNING</td>
<td></td>
<td>.32</td>
<td>.31</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td>.50</td>
<td>.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INITIATIVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ns means non-significant; r is greater than .05.
## INTERCORRELATION OF DIMENSION RATINGS

**TABLE 19:**
Intercorrelation of ratings from Condition 2 (n= 60 sets of ratings)

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>IJ</th>
<th>CI</th>
<th>BM</th>
<th>MR</th>
<th>E</th>
<th>CUD</th>
<th>P</th>
<th>ID</th>
<th>SR</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFORMED</td>
<td>.41</td>
<td>ns</td>
<td>.51</td>
<td>-.33</td>
<td>.54</td>
<td>.59</td>
<td>ns</td>
<td>.30</td>
<td></td>
<td>.63</td>
</tr>
<tr>
<td>JUDGMENT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMMAND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.25</td>
<td>ns</td>
<td>ns</td>
<td>.27</td>
</tr>
<tr>
<td>INFLUENCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUILDING</td>
<td></td>
<td></td>
<td>ns</td>
<td>ns</td>
<td></td>
<td></td>
<td>.34</td>
<td>.30</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>MORALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MONITORING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
<td>ns</td>
<td>.73</td>
</tr>
<tr>
<td>FOR RESULTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPATHY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISCIPLINE</td>
<td>.85</td>
<td>-.28</td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLANNING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.24</td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>INITIATIVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** ns means non-significant; r is greater than .05.
### TABLE 20:
Intercorrelation of ratings from Condition 3 (n=60 sets of ratings)

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>IJ</th>
<th>CI</th>
<th>BM</th>
<th>MR</th>
<th>E</th>
<th>CUD</th>
<th>P</th>
<th>ID</th>
<th>SR</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFORMED</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.48</td>
</tr>
<tr>
<td>JUDGMENT</td>
<td>.48</td>
<td>ns</td>
<td>.58</td>
<td>ns</td>
<td>.44</td>
<td>.58</td>
<td>ns</td>
<td>ns</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>COMMAND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.17</td>
</tr>
<tr>
<td>INFLUENCE</td>
<td>.17</td>
<td>.50</td>
<td>-.34</td>
<td>.47</td>
<td>.51</td>
<td>.24</td>
<td>ns</td>
<td>.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUILDING</td>
<td>ns</td>
<td></td>
<td>.27</td>
<td>-.31</td>
<td>.28</td>
<td>ns</td>
<td>.44</td>
<td>ns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MORALE</td>
<td>ns</td>
<td>.27</td>
<td>-.31</td>
<td>.28</td>
<td>ns</td>
<td>.44</td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MONITORING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.36</td>
</tr>
<tr>
<td>FOR RESULTS</td>
<td>-.36</td>
<td>.62</td>
<td>.69</td>
<td>ns</td>
<td>ns</td>
<td>.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPATHY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.65</td>
</tr>
<tr>
<td>CONSCIENTIOUS USE OF DISCIPLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.83</td>
<td>ns</td>
<td>-.40</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>PLANNING</td>
<td>ns</td>
<td>ns</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERPERSONAL DIAGNOSIS</td>
<td>ns</td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENSE OF RESPONSIBILITY</td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INITIATIVE</td>
<td>ns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ns means non-significant; r is greater than .05.
APPENDIX H:  
Content Coding Categories and Rules

# OF WORDS
*total count not including: a; an; the

ACTION STATEMENTS (action of person rated not action or reaction of others)
*active, action verb / or passive statement indicating that something was actually done by person to the passive subject (e.g. "information was gathered")
*following "ability to", "willing to" by active statement of what is done
*unambiguous statement that action occurred or person caused change (e.g. "take charge" does NOT count; "take action", "make decisions" Do count)

SPECIFIC REFERENCE TO ACTUAL OBSERVED BEHAVIOR
* reference to action/behavior in film
*reference to mood, tone, part of story does NOT count without reference to behavior of ratee

REFERENCE TO OUTCOMES
*outcome itself need not be explicitly stated
*need clear indication that outcome referred to followed from something person rated does or thinks

NON-BEHAVIORAL STATEMENT OF RATING CONSTRUCT
*global statement fo construct--no behavior
*phrases such as "ability to", "willing", "function of", "how" "can one", etc. followed by unobservable statements of intention or approach (e.g. "how person decided" does NOT count; "how person used resources" does count)
APPENDIX I:
Examples of coded passages

1

ACTION STATEMENTS
Passages coded:
♦ information gathering from experienced personnel
• I looked at how each member asked for opinions
• communication of standards and expectations through publicizing them

Passages not coded:
• how the men responded
• performance of squadron

SPECIFIC REFERENCE TO ACTUAL OBSERVED BEHAVIOR
Passages coded:
• interview with others in command
• Davenport collected all facts for the failed bombing mission and also questioned the 9000 ft. order.
• Gen. Savage enforced some regulations that had been unenforced for a long time.

Passages not coded:
• willingness to implement new policies which change an old system
• methods of gathering information and the effective interpretation

REFERENCE TO OUTCOMES
Passages coded:
• He was the major reason for the "hard luck" attitude because that was how he felt.
• He put together the Leper Colony to get certain people's attention.
• whether it was his action that actually sparked the behavior of others

Passages not coded:
• the way others responded
• getting the job done

NON-BEHAVIORAL STATEMENTS OF CONSTRUCT
Passages coded:
• being a leader
• the amount of control he seemed to have over his troops
• how well they deal with the unexpected when things go astray
• Initiative is the ability to get the ball rolling.

Passages not coded:
(phrases such as those coded under ACTION STATEMENTS or SPECIFIC REFERENCE TO ACTUAL BEHAVIOR)
APPENDIX J

Comparison of scores from condition 3 and 4:
Mean dimension scores and t-values

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean 1 (n=60)</th>
<th>Mean 2 (n=58)</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informed Judgment</td>
<td>1.4</td>
<td>1.7</td>
<td>-1.40</td>
</tr>
<tr>
<td>Command Influence</td>
<td>1.9</td>
<td>2.1</td>
<td>-.46</td>
</tr>
<tr>
<td>Influence</td>
<td>1.6</td>
<td>1.5</td>
<td>.55</td>
</tr>
<tr>
<td>Monitoring For Results</td>
<td>1.2</td>
<td>1.0</td>
<td>.79</td>
</tr>
<tr>
<td>Empathy</td>
<td>1.2</td>
<td>1.2</td>
<td>.41</td>
</tr>
<tr>
<td>Conscientious Use of Discipline</td>
<td>1.4</td>
<td>1.5</td>
<td>-.54</td>
</tr>
<tr>
<td>Planning</td>
<td>2.0</td>
<td>2.4</td>
<td>-1.05</td>
</tr>
<tr>
<td>Interpersonal Diagnosis</td>
<td>1.9</td>
<td>2.0</td>
<td>-.50</td>
</tr>
<tr>
<td>Sense of Responsibility</td>
<td>1.7</td>
<td>1.5</td>
<td>1.01</td>
</tr>
<tr>
<td>Initiative</td>
<td>1.5</td>
<td>1.5</td>
<td>-.13</td>
</tr>
</tbody>
</table>