INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand 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 you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in “sectioning” the material. 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 any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University Microfilms International
300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND
WARMKE, DENNIS LEE

EFFECTS OF ACCOUNTABILITY PROCEDURES UPON THE UTILITY OF PEER RATINGS OF PRESENT PERFORMANCE

The Ohio State University Ph.D. 1979

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

18 Bedford Row, London WC1R 4EJ, England
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 □
2. Colored illustrations □
3. Photographs with dark background □
4. Illustrations are poor copy □
5. Print shows through as there is text on both sides of page □
6. Indistinct, broken or small print on several pages □ throughout □
7. Tightly bound copy with print lost in spine □
8. Computer printout pages with indistinct print ✓
9. Page(s) □ lacking when material received, and not available from school or author □
10. Page(s) □ seem to be missing in numbering only as text follows □
11. Poor carbon copy □
12. Not original copy, several pages with blurred type □
13. Appendix pages are poor copy □
14. Original copy with light type □
15. Curling and wrinkled pages □
16. Other □
EFFECTS OF ACCOUNTABILITY PROCEDURES UPON THE
UTILITY OF PEER RATINGS OF
PRESENT PERFORMANCE

DISSERTATION

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

By

Dennis L. Warmke, B.A., M.A.

* * * * *

The Ohio State University
1979

Reading Committee:
Dr. Milton Hakel
Dr. Robert Billings
Dr. Leslie Berkes
Dr. Richard Klimoski

Approved By
Advisor
Dept. of Psychology
ACKNOWLEDGMENTS

I am very grateful for the help and support of many individuals. In particular I would like to thank the members of my reading committee, Milt Hakel, Les Berkes, Bob Billings and Randy Schuler and especially Rich Klimoski. Rich offered that perfect blend of leadership which provided structure when needed yet allowed the autonomy to research the questions I was truly most interested in. I am particularly thankful for the help Rich has given me in doing a refreshingly enjoyable dissertation.

I am also indebted to my friends at the Columbus Police Training academy for their genuinely supportive attitude toward research. It is rare to find a man with "Robbie's" (Captain Robinson's) interest in the self-renewal of his organization; Walt Rudin and Gary Miller were also tremendously helpful in their roles as trainers and chief counsels.

Finally, I would like to thank my wife, Janice for her patience and support. Of all the personnel selection decisions I may make during my career, Janice will always be the best.
I feel that there has to be someone else I have omitted from this list of acknowledgments, so I would like to thank my family (for introducing me to the joys of higher education, sound values and the game of tennis) and all of my friends and colleagues (for distracting me momentarily from schoolwork).

I have been lucky to get through ten years of higher education without losing my marbles and/or sense of humor. I am also lucky to have been aided by a dedicated and patient person like Mrs. Flora Cooney in the typing of this manuscript. Without her, my marbles and sense of humor might easily have been lost.
VITA

November 28, 1950 . . . Born - San Juan, Puerto Rico
1974 - 1978 . . . . Teaching Assistant & Teaching Associate, Department of Psychology, The Ohio State University, Columbus, Ohio
1977. . . . . . . . M.A., The Ohio State University
1979. . . . . . . . Visiting Lecturer, Department of Psychology, Denison University, Granville, Ohio

PUBLICATIONS


FIELDS OF STUDY

Major Field: Industrial and Organizational Psychology
Minor Field: Counseling Psychology
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>VITA</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>OVERVIEW</td>
<td>1</td>
</tr>
<tr>
<td>The Concept of Accountability</td>
<td>7</td>
</tr>
<tr>
<td>Hybrid Theory</td>
<td>16</td>
</tr>
<tr>
<td>Intention of the Study</td>
<td>24</td>
</tr>
<tr>
<td>The Issue of Generalizability</td>
<td>25</td>
</tr>
<tr>
<td>HYPOTHESES</td>
<td>28</td>
</tr>
<tr>
<td>II. METHODS</td>
<td>37</td>
</tr>
<tr>
<td>Participants</td>
<td>37</td>
</tr>
<tr>
<td>Experimental Setting</td>
<td>38</td>
</tr>
<tr>
<td>Experimental Procedures</td>
<td>39</td>
</tr>
<tr>
<td>Experimental Groups</td>
<td>43</td>
</tr>
<tr>
<td>Rating Scale</td>
<td>49</td>
</tr>
<tr>
<td>Reactions Survey Questionnaire</td>
<td>51</td>
</tr>
<tr>
<td>Measures of Rating Quality</td>
<td>52</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>58</td>
</tr>
<tr>
<td>III. RESULTS</td>
<td>60</td>
</tr>
<tr>
<td>Manipulation Checks</td>
<td>60</td>
</tr>
<tr>
<td>Effects of Procedures on &quot;Visible&quot; Measures</td>
<td>63</td>
</tr>
<tr>
<td>Effects on Interrater Agreement</td>
<td>68</td>
</tr>
<tr>
<td>Effects on Accuracy</td>
<td>72</td>
</tr>
<tr>
<td>Effects on Rater Reactions</td>
<td>76</td>
</tr>
<tr>
<td>Relations Among Quality and Reactions Measures</td>
<td>84</td>
</tr>
<tr>
<td>Acquaintance Survey</td>
<td>88</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>IV. DISCUSSION</td>
<td>93</td>
</tr>
<tr>
<td>Summary of Major Findings</td>
<td>93</td>
</tr>
<tr>
<td>Unexpected Results and Potential Explanations</td>
<td>98</td>
</tr>
<tr>
<td>Potential Limitations of the Present Study</td>
<td>107</td>
</tr>
<tr>
<td>Conclusions and Implications for Future Research</td>
<td>113</td>
</tr>
<tr>
<td>APPENDIXES</td>
<td></td>
</tr>
<tr>
<td>A. Landy and Farr Outline</td>
<td>123</td>
</tr>
<tr>
<td>B. Acquaintance Survey</td>
<td>125</td>
</tr>
<tr>
<td>C. Instructions by CEs</td>
<td>130</td>
</tr>
<tr>
<td>D. Captain's Comments</td>
<td>136</td>
</tr>
<tr>
<td>E. Checklist of Points to Remember when Making Evaluations</td>
<td>138</td>
</tr>
<tr>
<td>F. Rating Scale</td>
<td>140</td>
</tr>
<tr>
<td>G. Reactions Survey</td>
<td>148</td>
</tr>
<tr>
<td>H. Individual Reactions Survey Item Means Across Experimental Groups</td>
<td>154</td>
</tr>
<tr>
<td>I. Fortran Programs for Computing Dependent Variable Measures</td>
<td>159</td>
</tr>
<tr>
<td>J. Description of Columbus Police Department and Academy</td>
<td>165</td>
</tr>
<tr>
<td>K. Norms on Which Physical Performance Objective Scores Were Based</td>
<td>170</td>
</tr>
<tr>
<td>REFERENCE NOTES</td>
<td>173</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>174</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table | Page
-----|------
1: Means and Standard Deviations of Manipulation Check Questionnaire Responses by Experimental Group | 61
2: Values & Significance Levels of Effects of Review and Known Standards Conditions on Manipulation Check Items | 62
3: Means and Standard Deviations of Psychometric Indicators of Ratings Quality | 65
4: Multivariate and Univariate ANOVA F-Values for Review, Knowledge of Standards and Interaction Effects on Six Measures of Rating Quality | 66
5: Means and Standard Deviations of Interrater Agreement Measures of Rating Quality | 69
6: Multivariate and Univariate ANOVA F-Values for Review, Knowledge of Standards and R x KS Interaction Effects Upon Four Measures of Interrater Agreement | 71
7: Means and Standard Deviations of Accuracy 1 (Physical Fitness) and Accuracy 2 (Mastery of Subject Matter) Measures | 73
8: Multivariate and Univariate F-Values for Review, Knowledge of Standards and R x KS Interaction Effects Upon Accuracy 1 and Accuracy 2 Measures | 74
9: Correlations Among "Reactions Survey" Constructs | 79
10: Means and Standard Deviations of "Reactions Survey" Constructs by Experimental Group | 81
11: Multivariate and Univariate ANOVA F-Values for Review, Knowledge of Standards and Interaction Effects on Reactions Survey Constructs | 83

vii
Table 12: Correlations Among Measures of Rating Quality (Behavior) and Reactions Survey Constructs (Attitudes)........................, ............................................................ 86
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Constructs and Corresponding Items From the Reactions Survey</td>
<td>53</td>
</tr>
<tr>
<td>2</td>
<td>Revised Constructs, Internal Consistency Estimates, and Corresponding Items from the &quot;Reactions Survey&quot;</td>
<td>78</td>
</tr>
</tbody>
</table>
INTRODUCTION

The problem of obtaining accurate and useful subjective ratings in organizations has existed for some time. Such ratings can significantly influence organizational effectiveness by providing 1) information on which to base administrative decisions, 2) developmental feedback for employees, and 3) criterion data for test validation. Nonetheless, the quality of subjective ratings is often questioned, and their proneness to rater biases (sexual, racial) and errors (halo, leniency, etc.) is well documented (see Landy & Farr, 1979 for a review).

Past research has attempted to improve the quality of performance ratings in a number of ways. Though other areas have been investigated (Landy & Farr, 1979), the largest proportion of research has explored the effects of rating scale content and format upon rating errors. However, this approach has not been successful in explaining very much rating error variance (Borman & Dunnette, 1975; Schwab, Heneman & De Cotiis, 1975). Recently, more attention has been directed toward training raters to improve their rating and observational skills. This research has more consistently shown effects upon
the quality of experimental ratings (Bernardin & Walter 1977; Latham, Wexley & Pursell, 1975) but has not yet demonstrated generalizability to administrative ratings made in the organizational context (Spool, 1978; Warmke & Billings 1979). It has been suggested by Warmke and Billings that one reason why training may not generalize to administrative ratings is that it has typically been directed at familiarizing raters with the errors they may be prone to without dealing with organizational motivational contingencies which may not encourage rating quality.

One promising possibility for affecting rating quality within either experimental or administrative ratings is the theoretically based concept of accountability. Accountability provides motivation in a straightforward way: Person A can be said to be accountable to person B if person B has power or influence over person A and person A's behavior will be reviewed and or evaluated by person B. Under these conditions person A is typically motivated to comply with the expectations of person B. The concept has been studied primarily within negotiation contexts but has recently shown promise within performance rating contexts also (Decker, 1977; Goldmark, 1978). These studies have made raters accountable to some influential review source (e.g. their supervisors or peers) for the quality of their ratings.
The overall goal of the present study will be to assess the effects of accountability procedures upon the quality and utility of peer ratings. The selection of peers as a source of ratings for the present study was based primarily upon the promise these have shown for practical application within organizations. Peers often have better opportunities to observe performance relevant behavior than do supervisors, and peers also have a different perspective from which to assess performance (Borman, 1974; Klimoski & London, 1974, Lawler, 1967). Furthermore, considerable evidence exists demonstrating the predictive validity of peer assessments (see reviews by Kane & Lawler (1978) and Lewin & Zwany, 1976), and a recent study has shown markedly higher reliabilities, validity and resistance to bias within peer ratings of present performance than within supervisory ratings (Kisler & Nebeker, Note 3). Kane and Lawler (1978) have noted that assessments by peers "can contribute information that is psychometrically superior to or different from that which existing sources can contribute" (p. 555) and suggest that "the most imminent application of peer assessment on any appreciable scale seems likely in its adoption for performance appraisal purposes in work organizations" (p. 556). Though the present study will examine the effects of accountability procedures upon ratings made by peers only, the effects of accountability
processes should also be generalizable to ratings from alternate sources (e.g. supervisory, self, and subordinate ratings).

Unfortunately, little research exists dealing with factors influencing the quality or utility of peer evaluations of present performance, and the dynamics underlying such assessments are poorly understood. Therefore, the investigation of accountability effects upon the utility of peer ratings is important for the following reasons. First, it will help to expand our understanding of the accountability concept, its components, and its effects. Secondly, it will increase our present knowledge of the interrelations among the dependent measures typically used to measure rating quality or utility. Thirdly, it should have practical value to practitioners in demonstrating the utility of peer ratings within an operational appraisal system. The remainder of the present chapter will review the literature relevant to the concept of accountability and to efforts towards improving the quality of appraisals made by peers.

In an attempt to integrate and organize the rating research, Landy and Farr (1979) have proposed a model of ten major factors or independent variables influencing rating behavior. The model includes: 1) rater characteristics (e.g. sex, race, job experience), 2) ratee
characteristics (e.g. age, attractiveness, competence), 3) position characteristics (e.g. levels of rater and ratee in the organization, job difficulty), 4) the rating instrument (e.g. graphic, behaviorally anchored), 5) scale development procedures (e.g. use of developmental groups, item analyses), 6) rating process (e.g. who will see the ratings, rater training), 7) organizational characteristics (e.g. size, span of control, turnover), 8) data analysis procedures (e.g. statistically normalizing distributions, moderator analysis), 9) performance description (e.g. who the data is fed back to and how), and 10) personnel action (e.g. the effects on salaries, selection systems, etc.). An overview of the Landy-Farr model can be found in Appendix A. A literature review accompanying the model suggests that excessive research effort has focused on the effects of rating instruments (factor 4) and a moratorium on such research is called for. It is further suggested that attention should be shifted toward other variables in the rating model, such as rating process variables (factor 6).

Landy and Farr note that the term "rating process" includes the constraints placed upon the rater by requests and demands (in press). These requests and demands may be conveyed in various ways, including rating instructions which convey expectations, procedures, and consequences of ratings. For example, ratings may be
collected at varying time intervals, under varying conditions and time constraints, for varying purposes (experimental, administrative), with variations in who they must be justified to, and with variations in the clarity of conveyed expectations to raters (e.g. training, forewarning). Surprisingly, very little research attention has been directed at these important rating process factors; procedures and instructions accompanying the collection of ratings have often been haphazard, variable, and/or arbitrarily chosen. Furthermore, very little consensus on optimal procedures or processes exists within the supervisory rating literature.

Kane and Lawler (1978) have reviewed the peer assessment literature and note that the vast majority of reported studies were carried out under "research use only" instructions. Wide variations have also existed in the conveyed purposes of peer appraisals and the extent to which raters were identified and/or responsible for justifying their ratings. The effects of these process variations within supervisory or peer ratings have seldom been explored, and few attempts to explain these rating process phenomena have been made.

It would appear that a potentially powerful explanation for many of the process or demand variables noted above (e.g. who ratings must be justified to) may be found in the concept of accountability. The concept has
been studied primarily within negotiation contexts in which a negotiator must justify the agreement he has reached to his constituency. Knouse (1977) has reviewed the literature on accountability and defined the concept as:

the influence, both internal and external, upon an individual exerted by a source for the purpose of maintaining adherence to its sent expectations.

The internal dimension of accountability is concerned with the identification or internalization of goals of the influence source. The external dimension deals with the pressure to comply with the desires of the influence source (Knouse, 1977, Ford, Note 2).

Within the negotiation context, accountability has most often been operationalized by some sort of review by a constituency of a negotiator's performance. Other manipulations have altered the visibility of the reviewers and/or the clarity of their expectations. (see Klimoski, 1972; Haccoun & Klimoski, 1975; Organ, 1971). It has generally been found that the anticipation of review and greater constituency visibility result in the negotiator being more persistent in obtaining a decision favorable to the group (Klimoski, 1972) and in more adherence by the individual to the desires of the influence source (Knouse, 1977).

Though some theoretical development of accountability notions exists within the negotiation literature,
the application of the accountability concept within the performance appraisal domain has been rare. Apparently, only Goldmark (1977) has sought to explain the results he obtained via accountability reasoning. He used a manipulation which caused some raters to expect that their ratings would be reviewed by and discussed with other raters and found that this expectation increased the accuracy of ratings.

Despite the absence of research within the "rating process" and procedures domain, major calls for this type of research have been made by Wherry (Note 5) and Landy and Farr (1979). One theoretical formulation related to rater process and accountability issues is by Wherry (Note 5). Wherry's general psychometric theory of the rating process is founded on the premise that rating accuracy is dependent on 1) the performance of the ratee, 2) the observation or perception of the ratee's performance by the rater, and 3) the recall of the observations of performance by the rater. A total of 46 theorems and 24 corollaries concerning rating validity and reliability are presented. Among Wherry's theorems is one in particular which is directly relevant to rating process and accountability concepts. Theorem 12 states:

Knowledge that the rating given will have to be justified may serve unconsciously to affect the rating given. (p. 14).
The theorem is accompanied by three corollaries which state:

Corollary 12a. Knowledge that the rating may have to be justified to the ratee may cause the rater to recall a higher proportion of favorable perceptions and thus lead to leniency.

Corollary 12b. Knowledge that the rating may have to be justified to the rater's superior may cause the rater to recall a higher proportion of perceptions related to actions known to be of particular interest to the superior, whether such actions are pertinent or not.

Corollary 12c. To assure that neither of the above distorting affects shall take place alone, it is better to assure their mutual cancellation by requiring that both types of review shall take place.

Landy and Farr also call for more research relevant to rating process variables and the testing of Wherry's propositions. They note that "the administrative implications of Wherry's model of rating variance are important" (p. 67) and suggest that "the theory should be carefully examined by performance rating researchers and tested empirically to a much greater extent than has already occurred" (p. 66). It appears then that rating process variables related to rater accountability are important and merit greater investigation.

The theoretical reasons underlying the effects of accountability upon behavior are not clear. However, several plausible explanations can be suggested. One suggested by Wherry is that the need to justify ratings to others (one operationalization of accountability)
leads to unconscious distortions in the recall of behavior. As suggested by his Corollaries (12a-c), the accountability or justification source can be expected to differentially affect rating accuracy so that justification to ratees may decrease quality and justification to supervisors may increase quality.

A second theoretical explanation, suggested by Knouse, is social facilitation theory (Zajonc, 1969). This theory holds that the social presence of others increases a person's activation level and results in superior performance on well-learned or dominant responses, while decreasing performance on new or non-dominant responses. This theory would actually suggest that accountability to others could increase or decrease quality, depending on whether the rating task and/or the ratees were very familiar to the rater.

A third explanation of accountability effects can be found in classical behaviorist theory, as formulated by Thorndike and Skinner. Within this paradigm all behavior is held to be the result of past behavioral consequences (reinforcement history). This theory would hold that the consequences of past accountability experiences (whether positive or negative) will determine future rating behavior under conditions of similar accountability.
A fourth theoretical explanation can be broadly called cognitive theory. Cognitive theories of behavior are based on the premise that persons not only act on the basis of past experience (behavioral theory) but that the primary determinants of behavior are cognitions about the consequences of behavior. This theory differs somewhat from a behavioral theory in its predictions. It suggests that if a person's thoughts can be altered, that behavior will also change.

An outgrowth of cognitive theory has been called expectancy or VIE theory (see Mitchell, 1974). This theory holds that behavior will be determined by expectancy perceptions (the expectation that effort will lead to the successful performance of some action, i.e. accurate ratings), by instrumentality perceptions (the perceived relationship between successful performance and the consequences of that performance) and by the valences or values attached to the perceived consequences. This theory would suggest that potential raters will weigh the costs and benefits of alternative levels of effort and patterns of ratings and will choose to act in a way which maximizes benefits. This theory suggests that raters will consciously ask themselves "what consequences will different levels of effort and accuracy have for me?" and act accordingly.

A sixth explanation can be seen as closely related to cognitive expectancy theory and involves the fear of
sanctions or expectation of rewards as the explanation for accountability effects. Kogan, Lamm & Trommsdorf (1972) have suggested that the fear of sanctions is the major force influencing the behavior of subordinate negotiators. Cartwright (1965) also incorporates the notion of fear of sanctions and desire for rewards as primary determinants of social influence. This explanation, along with the cognitive and expectancy theories, place primary emphasis on the expected consequences of particular actions by the target person. Thus, the expectations of the accountability source (e.g. ratees, fellow raters, or superiors) would be seen as influential to the extent that these sources controlled important sanctions or rewards.

A seventh explanation of accountability effects involves face saving. This explanation, suggested by Knouse (1977), postulates that in order to avoid embarrassment a person will sacrifice considerable gain (also see Goffman, 1959, Brown & Garland, 1971). Organ (1971) has indicated that within negotiation contexts accountability may arouse the need to save face and involve the perception of the constituency as allocators of rewards and punishments for levels of adherence. Within a performance appraisal context, the face saving explanation would imply that raters would attempt to make themselves appear competent as evaluators and avoid appearing
incompetent. It also suggests that some gain would be sacrificed in order to achieve this goal (e.g. time spent, interpersonal conflicts resulting from accurate ratings).

A closely related explanation is interpersonal comparison theory, as presented by Myers and Lamm (1976). They have indicated that the concept involves "a type of normative social influence (connoting social-emotional processes and concern for favorable self perception and self presentation)." The authors note that interpersonal comparison theory suggests that the mere exposure to the preferences or expectations of others is the necessary and sufficient condition for behavior change. Thus, knowing the expectations of the source of accountability would be sufficient for ratings to be altered.

The eight explanations outlined above are certainly overlapping but generally not contradictory. They hold certain elements in common and differ in certain ways also. One way in which the theories differ is in their relative emphasis on conscious versus unconscious behavior determination. The first three theories, Wherry's, social facilitation theory, and behavioral theory, downplay the importance of consciousness. All three suggest that the behavior which can be expected will result with or without conscious decision processes. The remaining theories suggest a more active or conscious awareness of expectations, and behavior which is based on these. This
conscious-unconscious distinction among the theories appears to make little difference in the predictions that would be made by the theories. It may have implications, however, for the feasibility of measuring perceptions which respondents may or may not be aware of.

The theories have some important elements in common. Perhaps the most dramatic commonality is that all of the theories suggest motivation as the prime determinant of the target behavior. This makes intuitive sense but will be briefly explored. Any behavior can be seen as a function of capability (can) and motivation (will). Lawler (1973) and Campbell and Pritchard (1976) have outlined the basis of the Performance = (capability or ability) x (motivation) notion. Within the performance appraisal context the factors influencing performance (defined as making accurate ratings) can be seen to include capability or ability (which includes knowledge of both the behavior of the particular ratee and of the criterion behaviors, as well as knowledge of rating errors and individual rating aptitude) and the motivation to make accurate ratings. Individual rating aptitudes are largely not alterable, and knowledge of the ratee's criterion relevant behavior is not very easy to influence. However, rater knowledge about potential rating errors and the motivation of the rater to make accurate or inaccurate ratings would seem to be both important and controllable.
Therefore, it appears logical that a rater's motivation will be influential in affecting ratings.

It may be fruitful to explore commonalities among the abovementioned explanations of how accountability affects rater motivation. Perhaps the most distinctive commonalities among the eight theories noted above is that all have in common **energizing** and **directional** components or dimensions (Klimoski, Note 4). An accountability condition first causes the organism to prepare to alter its present behavior. Social facilitation theory, for example, suggests a form of tension brought about by the social presence of others, and the cognitive theories suggest that the anticipation of social encounters should create a similar tension which prepares the organism for action. All of the theories also suggest that the outcomes or anticipated outcomes of the accountability manipulation are important and will influence the direction of the rater's behavior. Wherry's theory, for example, suggested different predictions about the direction of ratings, depending on the accountability source (i.e. whether ratings would be justified to ratees or to superiors). This theory is consistent with the cognitive, VIE, and sanctions-rewards theories to the extent that raters and superiors would be expected to react differently to accurate ratings. Therefore, the explanations of how accountability affects ratings all involve energizing
and directional dimensions.

Though the predictions by the eight theoretical explanations of accountability may be quite similar, it may be advantageous to select one preferred theory around which to work. The theory to be employed here can be seen as a hybrid social comparison-expectancy theory. This will hold that individuals in general are concerned about favorable self perception and self presentation, and that they often look to the judgements of others as indexes of their own correctness (see Festinger, 1954; Myers & Lamm, 1976). This being the case, an individual will experience some arousal when it is learned that his or her judgements will be compared to those of others. In addition, the direction of one's own judgements are hypothesized to be influenced by his perceptions of the expectations of the comparison individuals, especially when no other standard of "correctness" exists. Furthermore, several characteristics of the comparison individuals, especially their capability to apply rewards and sanctions, will be influential. Other conditions, too, may serve to facilitate the effects of accountability procedures. The social-psychological conformity literature notes that conformity to the judgements of others increases if the task is ambiguous, and if the individual has low general or task specific self esteem (see Aronson, 1972). This literature also indicates that the
characteristics of the comparison group (i.e. those who will review one's judgements) can be influential. The rater's similarity or attraction to the comparison individuals can increase their influence. Finally, it is logical to believe that the extent to which accountability procedures can affect motivation and rating quality must depend somewhat on the initial motivation of the rater to make the evaluations. It is also likely that accountability will have greater effects on ratings on which there is low initial motivation to rate accurately because increased motivation will then have more potential for bringing about behavior change. In summary, the social comparison-expectancy theory being proposed suggests that rating quality is influenced by two major components of accountability: 1) review by an influential source (e.g. telling a rater that his/her ratings will be compared to and/or reviewed by a source capable of administering rewards or sanctions, and 2) Knowledge of the standards to be used by the review source.. (i.e. telling a rater specifically what will be expected of his/her ratings).

Empirical research relevant to the above theory can be derived from two major sources: the accountability related literature and the rater training literature. The accountability literature has generally varied the accountability source (whether ratings would be justified to
ratees, superiors, or fellow raters) without directly manipulating the expectations or standards of the accountability source. An exception to this was the Decker (1977) study, which manipulated accountability to superiors and also attempted to manipulate the perceptions of raters about the expectations of their superiors. The remaining accountability related studies have apparently assumed that there would be considerable agreement among raters in their perceptions of the expectations and standards of the accountability source.

Little research to date has manipulated rater accountability to ratees or to their own superiors. Sharon and Bartlett (1969) included a justification condition in which the rater identified himself and perceived that he would have to justify, in a face-to-face situation, his evaluations to ratees, and it was found that accountability to ratees increased leniency error. Stockford and Bissel (1949) also found that raters who perceived that ratings would need to be explained to subordinates rated in a more lenient manner.

Decker (1977) found that raters who were accountable to their own supervisors rated more leniently. However, there were obvious problems with the study. For example, Decker attempted to manipulate the expectations that the superiors conveyed to raters with respect to rating accuracy; however, manipulation checks showed
that this manipulation was not effective and that raters perceived about the same level of expectations of accuracy from superiors. In addition, a post-experimental questionnaire suggested that raters perceived more lenient ratings in this situation as serving to increase their self-esteem vis-a-vis the superiors. This study, far from contradicting the social comparison-expectancy theory, indicates the importance of making clear the expectations and/or standards of the accountability source if higher quality ratings are to be expected.

Perceived accountability to fellow raters has been used frequently in practice, but only one study to date has directly manipulated this kind of accountability. Goldmark (1978) found that raters who expected to have to discuss their ratings with fellow raters made more "accurate" ratings in the differential accuracy sense.

Goldmark (1978) used college students as subjects within a laboratory setting. Videotapes of a salesperson served as the standard stimulus. One independent variable altered level and timing of accountability. Three conditions were used: In the first, no accountability (anticipation of discussing ratings with fellow raters) was projected; in the second, raters were told that they would be discussing their ratings with fellow raters only after they had finished observing the videotapes; and in the third, raters were informed of later
discussions with fellow raters prior to observing the videotapes. The results of the study showed that both accountability conditions (early and late expectation of discussion) increased rating accuracy compared to the no expectation of discussion condition.

Goldmark's study can be seen as strongly supportive of social comparison-expectancy theory, because it contained both major ingredients considered necessary for accountability effects. One important difference from the Decker study was that Goldmark made very clear to raters what was expected in terms of rating quality. Raters were given preliminary training in how to make accurate ratings, raters knew that the persons they would discuss ratings with had been exposed to the same material (and were thus equally expert raters), and the standard for accuracy was rather unambiguous because the ratees' behavior was available as a criterion on the videotapes. Thus, raters knew their ratings would be compared to others, had clear expectations of and a standard for rating quality, and the comparison group was composed of fellow college freshmen very similar to the raters.

In addition to the experiments viewed above, several training-related studies have included accountability manipulations which may partially explain their success. One case study by Kissler and Nebeker (Note 3) suggests support for the effectiveness of accountability
to fellow raters accompanied by some form of training or instructions which sets a "standard" for rating quality. Subjects in this study were research scientists employed by the federal government, who made peer ratings within an appraisal system to be used for implementing ongoing promotion decisions. Within this system raters were responsible for making peer ratings and for meeting and discussing these ratings with fellow raters. Kissler and Nebeker (Note 3) found outstandingly high quality ratings resulting from these conditions. Interrater agreement (prior to discussion) was above .9, reliability over time (test-retest over 3-year intervals) was about .9, and correlations of ratings with objective indexes of research productivity were approximately .5. These ratings were considerably superior to supervisory ratings of the same individuals on several criteria of rating quality. The high quality of these peer ratings may have been due to the highly structured system used, which the authors suggest may have served as a type of rater training, and/or to the accountability of raters to fellow raters who were very much like themselves. Kissler and Nebeker note that "the anticipation of discussion may have resulted in a more thorough performance evaluation" (p. 15).

Several other studies dealing with rater training tend to support the potential importance of accountability.
It is quite interesting to note that some of the most effective training programs have made rater accountability a part of the training. One of the most often referenced (and best executed) training programs to date, by Latham, Wexley, and Pursell (1975), incorporated rater accountability to fellow raters as one of its ingredients. The 9-hour "intensive workshop" reduced several rating errors, and included conditions in which ratings by trainees were reviewed by the trainer and evaluated for accuracy. Similarly, Borman and Rosse (Note 1) who used a shortened 3-hour version of the Latham et al. workshop included time in which ratings by each participant were placed on a flip chart and "differences between ratings and reasons for these differences are argued by participants in a discussion led by the trainer" (p. 10). This training also proved effective in reducing halo error across five different rating scale formats. Though these training studies contained many other possible ingredients that might account for their success (e.g. lecture on rating errors, practice, feedback) it is significant that review by trainers or fellow raters also was included. These findings, along with those by Goldmark and Kissler and Nebeker, add solidarity to the hypothesis that accountability to fellow raters improves rating quality.
These empirical studies lend some support to the notion that different accountability sources may consistently lead to different perceptions on the part of raters in terms of expectations about rating quality. These studies also suggest that training serves to clarify standards of rating quality and rater expectations, and these ingredients may be necessary for accountability manipulations to have effects.

A final body of research which suggests that accountability is correlated with higher quality ratings is the assessment center literature. Within assessment centers, typical procedures call for assessors to view the behavior of ratees and to independently rate their performance (see Finkle, 1976 for a review). Assessors typically meet in groups where they must discuss the ratings with fellow raters who have observed the same sample of behavior. Those assessment center studies that have evaluated rating quality have typically shown very high quality of ratings, particularly with respect to interrater agreement (Schmidt & Hill, 1977).

The assessment center situation conforms closely to the conditions suggested earlier in the social comparison-expectancy theory formulation. Overall, the empirical data suggest that some forms of accountability can powerfully influence the quality of ratings; they also suggest support for the utility of the social comparison-
expectancy theory of rater motivation.

In summary, a need for explaining and reducing rating errors exists. However, much of the past research has focused on rating instrument formats. It would appear that important strides can be taken if more attention is directed at rating process variables, such as accountability. Accountability has some theoretical underpinnings and is supported by scattered and indirect empirical evidence. Therefore, a study which more systematically examines accountability notions could contribute significantly to the rating literature.

The present study will test the utility of procedures for increasing accountability. The study will make a contribution for several reasons: First, it will provide a more definitive test of the effects of accountability manipulations than have been provided thus far. Only one true experiment (by Goldmark) has explored accountability effects on rating quality, and it has not provided solid evidence of accountability effects. Secondly, past research dealing with rating quality has been largely atheoretical. The present study has suggested a theoretical rationale of variables affecting rating quality. Thirdly, most research to date has for one reason or another been of questionable generalizability. The generalizability issue merits further discussion prior to the presentation of hypotheses.
The question of generalizability has often been addressed in the research methodology literature under the rubric of "external validity." Campbell and Stanley (1963), Cooke and Campbell (1976), and Runkel and McGarth (1975) note that external validity refers to the extent to which results of a study can be generalized across different subjects, experimental manipulations, dependent variables, and contexts. Within the rating literature little effort has been expended in dealing with these issues.

It is beyond the scope of the present study to document completely all potential threats to the external validity of past research. However, several examples will illustrate the case. First, the subjects in rating studies have most often been university students, who differed considerably from raters in actual organizations, both in terms of exposure to ratees and motivations to rate accurately. A second generalizability problem can be traced to an almost universal absence of "multiple operationism." Seldom have multiple dependent measures been used as indexes of rating quality, and rarely have different operationalizations of a given independent variable been employed. Thirdly, most rating studies have been carried out in contexts which differed importantly from operational rating conditions. For example, most rating studies collected rating data which was
anonymous (as opposed to identified), and which would be used for "experimental or research" purposes, instead of for making organizationally relevant decisions. Landy and Farr (1979) have clearly noted that both of these conditions have been shown to significantly influence rating quality. A related generalizability problem arises from the settings used. The carryover of results obtained in lab settings has seldom been crosschecked within actual organizational settings. Furthermore, reactivity, experimenter effects, demand characteristics, and Hawthorne effects may lead to differences by setting. Another problem concerns generalizability across rating instruments. Rating instruments used in research may differ significantly in format (e.g. nominations vs. ratings, graphic vs. behaviorally anchored) from those used operationally in organizations. A final problem concerns the absence of evidence of generalizability across rater sources (e.g. supervisors, peers, subordinates). Comparisons across sources in terms of rating quality have been infrequent (see Holzbach, 1978), and comparisons which have provided realistic conditions, and treated the sources in comparable ways, are almost nonexistent (see Kissler & Nebeker, Note 3, for a near exception).

Some of these generalizability problems have been pointed out within the peer assessment literature by Kane
and Lawler (1978). These authors have noted that most peer assessment studies have used peer nominations for the purpose of predicting future performance, and these nominations have generally been collected under a "research purposes only" set. The generalizability of these findings to peer ratings of present performance to be used for making organizational decisions remains almost entirely unexplored.

Problems with generalizing the results of past rater training studies have been noted by Warmke and Billings (1979). They suggested that the major threat to generalizability may be the operation of organizational-contextual demands and constraints upon raters which take effect only when ratings are to be used for organizationally relevant purposes. It was suggested that these contextual motivational factors may be more important in influencing rating quality than rater training.

The problems outlined above suggest that if the results of rating studies are to be generalized to organizational rating conditions, and are to have direct utility, efforts should be made to deal with known threats to external validity. In order to deal with such threats the present study will attempt to 1) employ a field site or sites, 2) use rating scales and formats capable of being used for operational decision making, 3) employ
ratings to be used for organizationally relevant purposes, 4) use multiple measures of rating quality, and 5) employ independent variable manipulations which are feasible for operational use. These precautions should guard against many potential threats to external validity and greatly enhance the potential generalizability of obtained results.

HYPOTHESES:

Effects of accountability procedures upon compliance with rating quality standards. The literature on accountability has suggested that in order for maximum accountability to be manifested, two ingredients (knowledge of internal or external pressure [i.e. review], and knowledge of the accountability source's standards) should exist. These conditions, according to Kelman's (1961) sources of influence, will create maximum compliance with the accountability source's standards.

The literature cited earlier clearly indicates that a combination of knowledge of standards and review will generally lead to greater compliance than a comparable situation lacking in these (see Knouse, 1978 for a review). However, the literature has little to say about the manner in which these two components of accountability interact or the relative potency of the components.
With respect to the manner of interaction, the two components can be viewed as interacting either in a multiplicative or an additive way. A multiplicative explanation might argue that the knowledge of standards and review components are comparable conceptually to the ability x motivation components of Lawler's (1973) performance equation.

The imparting of knowledge of standards to be used in evaluating performance is analogous to conventional training in that it implies the relating of information about what constitutes accurate and/or inaccurate ratings. This training, in turn, should affect the ability or skill of the rater to successfully translate ratee behavior into evaluations. As was noted earlier, the review component of accountability is clearly intended to increase the external motivation of the rater to conform with suggested rating quality standards. Review serves to set up contingencies in which ratings which fail to meet the standards may be punished (either verbally or otherwise) by the reviewer. Thus, knowledge of standards may increase a rater's ability to evaluate more accurately and review may motivate the rater to fully apply his evaluative skills. Given a multiplicative interaction of knowledge of standards and review, it could be argued that if either component were totally absent, the performance on the rating task would be very
poor (i.e. random ratings). However, it is clear that in practice knowledge of standards is only one aspect of rater ability and review is only one aspect of motivation. Furthermore, neither ability nor motivation will be totally absent in actual practice within either a "no review" or a "general standards" condition. Nonetheless, the multiplicative interaction notion would argue that in practice a group with one accountability component (i.e. review) will perform better on the rating task than a group with neither component (control).

A similar prediction to that for the multiplicative model would be made within an additive interaction of components model. Again, the group with only one of the components would be expected to outperform a group with none. Finally, it should be clear that either a multiplicative or additive model will predict higher performance (greater compliance) for a group containing both components than one with only one.

Information on the relative potency of the two components is not available at this time. However, some support for the notion that knowledge of standards, without some mechanism for encouraging compliance, will function to alter ratings is provided by the rater training literature. Nonetheless, it should be noted that most rater training studies have attempted to alter only the characteristics of ratings made for experimental
rather than administrative purposes. To the extent that motivational forces differ greatly within these two contexts, it is possible that raters may choose not to abide by the standards presented during training. Unfortunately, no research to date has compared a review condition to a training condition (within any setting) so the relative potency of the two components remains unknown. For purposes of the present study no prediction will be made which would suggest a difference in potency between the two single component groups. In summary, there is evidence that the presence of both components will significantly increase compliance over a condition with neither component. Moreover, regardless of whether the components interact in a multiplicative or additive way, a condition with one of the components should outperform one with neither and underperform one with both. On this basis the following hypothesis is proposed:

Hypothesis 1: Compliance to rating quality standards will increase as the number of accountability components increases. Compliance should be highest for the known standards-review group, intermediate for the known standards only and review only groups, and lowest for the control group with no known standards and no review.

Effects of accountability procedures upon convergent validity measures of rating quality. Based upon both logical deduction and limited empirical evidence it would appear that experimental manipulations may affect some measures of rating quality but not others. The
accountability procedures used here may influence compliance with the rating quality standards (and thus measures of halo, leniency and variability) yet not influence the true validity or accuracy of the ratings. This prospect is of obvious practical as well as theoretical importance. Borman (1975), for example, showed that his experimental manipulation (a 5-minute lecture designed to impart knowledge of rating quality standards) was quite successful in leading to compliance with the suggested standard (pooled variance "halo" was reduced). However, Borman also found that the "accuracy" (convergence of ratings with true scores) remained unaffected and inter-rater agreement (another type of convergent validity) was reduced by the experimental manipulations.

Borman's results suggest that some measures which have been unquestioningly used as indicators of rating "errors" may be purposely alterable by raters without a corresponding increase in the validity or "accuracy" of the ratings. The present experiment will follow Borman's lead and explore not only the effects of accountability procedures on visible measures of rating quality (such as halo, variability, leniency) but also the effects upon accuracy and interrater agreement. Objective data will be collected to serve as a standard by which the "accuracy" of ratings on two performance dimensions may be evaluated.
The limited evidence to date, provided primarily by Borman (1975) suggests that compliance with rating quality standards may be altered within the present study but that convergent validity indexes may not be. The following hypothesis is proposed:

Hypothesis 2: Interrater agreement and "accuracy" (as indexed by the relation between ratings and objective performance measures) will not be significantly influenced by accountability procedures aimed at altering measures of halo, variability, leniency, or use of written comments.

Effects of accountability procedures upon participant reactions to the rating process. Since the research on peer assessment contains only anecdotal evidence regarding rater reactions to the process, and since the evidence seems to indicate moderate to high levels of resistance (see Kane and Lawler, 1978) the present study will investigate these reactions in a systematic way. The literature on accountability and compliance suggests strongly that review is viewed negatively by those who anticipate experiencing it. At least three reasons exist for this within a peer appraisal context. First, review places a threat of potential sanction imposition upon the rater. The rater knows that if he fails to comply, he may be punished. Second, the rater's perceived autonomy or freedom is reduced by review. Knowing the consequences of non-compliance, the rater may see little choice but to comply with the standards that are given (regardless of
the resulting accuracy of the ratings). Hackman and Oldham (1975) have noted that tasks low on the core dimension of autonomy are viewed as less motivating and satisfying. Third, conformity to the rating standards given requires that some negative information about peers be supplied and supplying this information may be distasteful to raters. Webb (1955) has indicated that raters resist having to give negative nominations in peer assessment settings. In the present study, the standards of distributing the ratees in terms of overall performance and of listing both positive and negative written feedback may create a similar resistance.

Knouse (1977) has commented that:

Review in this study appeared to be basically aversive with an underlying dynamic of evaluation apprehension (Klimoski, 1972; Rosenberg, 1969), p. 111.

Knouse further notes that "representatives anticipating final review wished to avoid negative evaluations," and review also tended to "increase the difficulty of negotiation--longer time to reach agreement and greater perception of pressure" (p. 111). Finally, he notes that one reason why his review group was relatively low in commitment may have been that "review as a hindrance to freedom may be interfering with volition to action" (p. 113).

Beyond the conclusion that review seems to decrease rater satisfaction and increase resistance to the process,
little empirical data exists. For example, no studies reviewed by the author have studied the effects of training upon rater reactions. Furthermore, no data on review-training interactions upon reactions can be found. Nonetheless, it is logical to argue that a review condition without knowledge of standards would be frustrating to a rater. Likewise, it may be reasonable to assume that knowledge of standards by itself may either reduce, increase or fail to influence task satisfaction. The most relevant literature may be in the leadership domain and deals with the influence of structuring activities upon subordinate satisfaction. The results seem to indicate that there is generally no consistent relationship between structuring and satisfaction (Fleishman & Harris, 1962). Therefore, no sound basis exists for predicting differences in reactions by raters in control and knowledge of standards only groups.

Overall, the data indicates that review will lead to negative rater reactions, and logic indicates that review without standards will be more aversive than review with known standards. The following hypothesis is proposed:

Hypothesis 3: Rater reactions to the rating process will be more negative when review is present than when no review is present. Resistance to the process should be highest for the review-no standards group, second highest for the review-known standards group, and lowest for both the known standards only
and no standards no review (control) groups.
METHODS

The methods chapter of the present report outlines the procedures each experimental group underwent. This chapter will include sections on Participants, Experimental Setting, Experimental Procedures, Experimental Groups, Rating Scale, Questionnaire Measure of Rater Reactions, Measures of Rating Quality, and Statistical Analyses.

Participants:

Participants in the experiment were 53 police recruits undergoing pre-employment training at the Columbus, Ohio Police Training Academy. Thirty-nine (74%) of the recruits were white and 14 (26%) were black. Thirty-eight (72%) were male and 15 (28%) were female. The high level of racial heterogeneity was largely the consequence of legal proceedings which resulted in court imposed minority hiring "goals" or "quotas" for the Police Department. All recruits had at least a high school education, and a few also had some university training.
Experimental Setting:

The training academy complex was located in a suburb on the west side of Columbus and consisted of 2 major buildings. The first building contained all administrative offices, a large lecture hall with capacity for seating 70, and an indoor firing range. The second building consisted of 2 "module" classrooms, each capable of accommodating 20 recruits.

Each recruit underwent a total of about 20 weeks of intensive training. The majority of training consisted of in-class lectures on patrol operations, legal issues and related topics; this was accompanied by outdoor activities which included physical training and firearms practice. The present class began training on December 2nd, 1978 and graduated on April 25th, 1979.

The academy staff consisted of one captain (the Training Bureau Commander), 1 lieutenant, 3 sergeants, and approximately 12 police officers (six of these were referred to as Counselor Evaluators [CEs] and were charged with in-class instruction and the counseling and evaluation of 7-10 recruits).

During the first week of training recruits were divided into two platoons. Each platoon consisted of three "CE" groups (of from 7-10 recruits) each led by a Counselor Evaluator. Assignment to CE groups was done for administrative convenience and was not entirely random.
CE groups differed in average height so that attractive drill formations would be facilitated. The six CE groups were assigned to do most outdoor activities together and thus members of these groups necessarily interacted with one another to a moderate extent.

Entrance into the training academy was obtained by the author through the cooperation of the Training Bureau Commander. Six months prior to the primary data collection sessions, the author had agreed to assist the academy in testing and evaluation procedures in exchange for research privileges with recruits.

Experimental Procedures:

On the second day of class (December 3rd), participants were informed by the author that they would later be asked to participate in a peer evaluation study as part of the author's research. Ten weeks into the training period (in late February) participants were asked to fill out an "Acquaintance Survey" by the author. This questionnaire can be found in Appendix B of the present report and was intended to determine whether recruits knew each other well enough to evaluate one another fairly. Recruits were not told of the intention of the survey, except that the information was to be used as part of the author's research, and that its results would be confidential and unavailable to the academy staff.
Two weeks prior to the main data collection portion of this investigation, two of the academy Counselor Evaluators (CEs) were personally contacted and asked for their help as co-experimenters. These two CEs (officers "R" and "M") were chosen on the basis of their educational backgrounds, their favorable reputations at the academy, and their knowledge of experimental design.

The nature of the study was explained to both CEs, and copies of the proposal for the present research were given to each. Two meetings were arranged with these CEs prior to the time of data collection and their input into experimental procedures was requested. The CEs helped, for example, to develop the behavioral anchors for two of the dimensions on which recruits would be evaluating each other. The likelihood of resistance to the peer rating process was also discussed and questions were answered.

On the morning of the day of data collection (March 9th, 1979) the CEs were given outlines of the basic information to present to the two experimental groups each would administer. These outlines are included in Appendix C of the present report.

A meeting was also held with the Training Bureau Commander and the Academy drill sergeant in which procedures for directing recruits to the designated classrooms at the designated times was discussed.
All recruits were to meet at 1 PM in the main lecture hall of the academy. The drill sergeant then read out the names of a group of 14 recruits who were told to report to "module" classroom #1, and a group of 13 recruits who were told to report to "module" classroom #2 from 1-3 PM. The remaining 26 recruits engaged in firearms practice until 3 PM, at which time they too were told to report to module classrooms 1 or 2.

Participants were assigned to the four experimental conditions (each will be described later) in a stratified random manner. This was based on the a-priori expectation that race, sex and CE group membership might correlate with academy performance levels. Furthermore, rating biases caused by these factors have been well documented (Landy & Farr, 1979). For these reasons participants were assigned to the four experimental conditions in a stratified random manner which insured that race, sex, and CE group composition was equalized across experimental groups to the greatest possible extent. Participant assignment to conditions was carried out as follows: Participants were first grouped into the six CE groups described earlier. White males were then randomly assigned from CE group 1 to the experimental groups, and white males were similarly assigned from each CE group to the four experimental groups. Likewise, black males, white females, and black females were in turn
assigned randomly from each CE group into experimental groups.

A stratified random assignment procedure was important because during the course of the study recruits would be evaluating only other members of their own CE group. Therefore, if there were differences in skill levels across CE groups and disproportionate representation of CE groups within experimental groups, confounding could occur. In addition, due to potential rater biases equal representation of race and sex groups was desired within experimental groups. The present procedures insured that the raters (and ratees) in each of the four experimental groups were randomly assigned to conditions and did not differ significantly in race, sex or CE group membership. These procedures resulted in experimental groups with 14, 13, 13, and 13 participants each.

Participants reported to the module classrooms on the day of the study itself without being told why they were being assigned there. Following arrival into the module classrooms the Training Bureau Commander informed recruits of the purpose of their presence, and requested that all recruits participate in the study and make high quality ratings. The outline of instructions presented by the captain to recruits can be found in Appendix D.
Following these preliminary comments by the captain, the CEs in charge of these two groups presented further instructions intended to vary perceptions of recruit accountability for peer ratings to be made. The content of these instructions is described next in a section on Experimental Groups.

At 3 PM, after the first two groups of recruits had finished making ratings, they were directed to meet outside the main academy building. As this was happening the remaining 26 recruits were told which module rooms they were to report to and were directed not to interact in any way with the previous two groups of recruits. The recruits with the 3-5 PM meeting times also reported to modules 1 and 2, were given introductory comments by the Training Bureau Commander, and were exposed to the experimental procedures appropriate for their particular group.

Experimental Groups:

Group 1 (High Specific Knowledge of Standards, and Identified Review—High Accountability). Fourteen recruits were assigned to Group 1 which was administered by CE "R" and met from 1-3 PM in Module Classroom #1. The outline of instructions presented by both CE "R" and "M" to this and the other experimental groups can be found in Appendix C.
Recruits in Group 1 (and also in the other 3 experimental groups) were first presented with a standard set of General Introductory Comments. These comments noted that a) the use of peer ratings was widespread, b) peer ratings tended to be valid because of the great exposure of peers to relevant performance-related behavior, c) the purpose of these ratings was primarily for feedback purposes, but the average ratings received by each recruit would become a part of their academy file and would be available to decision makers and supervisors. In addition, it was noted that the weight to be attached to these peer ratings was uncertain because peer ratings had not been used at the academy before, d) all recruits would be making peer evaluations; however, slightly different procedures were being followed with the four groups of recruits in order to find out which procedures were most effective, e) after the evaluations were completed they would be turned in to each recruit's CE who would summarize the ratings each recruit received (in anonymous form) and provide recruits with feedback regarding the ratings they received, f) recruits were told that the peers they evaluated would not know who specifically had evaluated them, g) recruits were instructed to remove and look at the contents of Rating Packets which had been handed to them. A rating Packet consisted of a 10 x 13 manilla envelope containing a Rating Form and a Reactions Survey
questionnaire. In the Identified Review groups (1 & 2) the rater's name was written on the manilla envelope and on the Rating Form inside. Each recruit was told to look at the Rating Form and to note the names of the recruits he/she would be evaluating. Each recruit evaluated nine fellow recruits (at least seven of these were members of the rater's own CE group, the remaining ratees were randomly selected from other CE groups). Examples of the rating dimensions were read aloud to recruits and each was told of the importance of giving a numerical rating to each ratee on each rating dimension. Recruits were also referred to a Reactions Survey in their packets and told that this survey was to be used for research purposes only.

Following these General Introductory Comments participants in Group 1 were presented with "Review Condition Instructions" (also see Appendix c). These instructions noted that in addition to the self-development purpose of these ratings, another purpose was to help CEs to assess how good each recruit was at observing and evaluating his/her fellow recruits. Thus, each set of ratings would be looked over by the recruit's own CE and evaluated to see that the ratings and feedback provided were accurate. Because raters in the review condition had their own names on the rating form, it was apparent to them that their CE would know they were
responsible for the ratings. Recruits were told that their ability to evaluate others was an important index of their own potential as a future supervisor, because one of the duties of a supervisor is to accurately evaluate others. Finally, recruits were reminded that the quality of the ratings and feedback they provided would be reviewed and evaluated individually, and recruits were asked to try to be as conscientious and accurate as possible.

Recruits in Group 1 were next presented with Training Condition instructions intended to provide them with high specific knowledge of the standards to be used in evaluating the accuracy of their ratings. A handout was distributed which noted each of the standards to be used for evaluating rating accuracy. This handout was entitled "Checklist of Points to Remember When Making Evaluations" and can be found in Appendix E. Four major points were noted on this handout: 1) Good ratings should behaviorally describe typical present performance, and suggest ways for recruits to improve, 2) Good ratings should reflect the real differences among recruits in overall performance, 3) Good ratings should realistically point out individual's own personal strengths and weaknesses, and 4) Raters should try to base ratings on performance rather than on who they like or feel similar to. The first three major points on this outline were
intended to correspond to measures of rating quality used as dependent variables in this study. These measures will be described later in the Measures of Rating Quality portion of this chapter. The "Checklist of Points" handout also noted specific behaviors which an evaluator could perform which would satisfy the standards for rating quality. Recruits were told to read this handout, and the CE administering this condition read at least the major points on the handout aloud.

Following the Training Condition instructions recruits in Group 1 were given a set of Final General Instructions. They were told 1) to read the instructions on the rating form and to try to follow these, 2) they would have 80 minutes to complete their ratings of 9 of their fellow recruits, 3) they should spend no more than 2 minutes or so per page on the 36 page rating form, 4) to be sure to complete the entire rating form by 2:30, 5) to try to make their ratings as accurately and objectively as possible, and 6) that if there were any questions about the Rating Form or Reactions Survey that they should raise their hand.

Group 2 (Low General Knowledge of Standards and Identified Review--Moderate Accountability). Thirteen recruits were assigned to this group which was administered by CE "R" and met from 3-5 PM in Module Classroom #1. Group 2 underwent experimental procedures
very similar to those of Group 1. The only major difference was that Group 2 did not receive the "Training Group" instructions received by Group 1 which were intended to clarify the standards to be used by CEs for evaluating rating accuracy. Group 2 was thus presented with the same General Introductory Comments, Review Condition instructions, and Final General Instructions as was Group 1, and in the same order.

Group 3 (High Specific Knowledge of Standards, and No Identified Review--Moderate Accountability). Thirteen recruits were assigned to this group which was administered by CE "M" from 3-5 PM in Module Classroom #2. Group 3 participants were presented about the same instructions as those in Group 1 except that no specific Review Condition instructions were given to Group 3. When subjects in this group were given their General Introductory Comments they are told to remove their names from the covers of the rating packets (names were written on masking tape affixed to the packets), so that ratings could be made anonymously. Recruits were told that even their own CE (to whom the ratings would be turned in) would not know which ratings they had made. Unlike subjects in Group 1 and 2, subjects in Group 3 were not identified, and the quality of their ratings could not be reviewed since their names did not appear on either the rating packet or Rating Form.
Aside from the removal of their names from the rating packets recruits in Group 3 received the same General Introductory Comments, Training Condition instructions, and Final General Instructions (but not Review Condition instructions) as recruits in Group 1.

Group 4 (Low General Knowledge of Standards and No Identified Review--Low Accountability). Thirteen recruits were assigned to this group which was administered by CE "M" from 1-3 PM in Module Classroom #2. Recruits in Group 4 were given the same instructions as those in Group 3 except that no Training Condition instructions were given. Therefore, these recruits received only the General Introductory Comments and Final General Instructions.

Rating Scale:

The rating scale used in the present study was a considerably modified version of the behaviorally anchored rating scales (BARS) for police officers developed by Landy, Farr, Sall and Freytag (1976). The present scale consisted of the following four performance dimensions 1) Physical Fitness, 2) Compatibility, 3)
Mastery of Subject Matter and 4) Leadership. Each dimension was anchored with behavioral descriptions of high (9), average (5), and low (1) levels of performance. A copy of the Rating Form used (including the behavioral anchors) can be found in Appendix F. Each Rating Form contained enough pages to evaluate 9 fellow recruits on the four performance dimensions.

The selection of the four dimensions used was based on the following considerations. First, these dimensions have considerable face validity and are important to academy performance. Second, these dimensions are conceptually independent. For example, physical skills and academic skills (dimensions 1 and 3) are known to be relatively independent, and compatibility and leadership (as dimensions of personality) should not be highly related to either physical or intellectual capabilities. Third, we desired dimensions which could be anchored behaviorally, and fourth, we wanted to select at least two dimensions for which independent objective measures were available for cross validation. As mentioned earlier, for both the Physical Fitness and Mastery of Subject Matter dimensions such objective indexes were available from academy files. In-class academic exam scores were available for all recruits as an index of Mastery of Subject Matter, and objective scores on a standardized composite of six objective physical
performance measures (including time in 1 1/2 mile run, and number of pushups and situps) were available as an overall index of Physical Fitness.

The behavioral anchors for two of the four dimensions (Compatibility, Leadership) were drawn directly from the rating form in current use at the academy. The anchors for the Mastery of Subject Matter and Physical Fitness dimensions were developed through discussions with academy CEs. Overall, it was felt that the Rating Form dimensions used captured a number of relatively independent and important aspects of police recruit performance.

Reactions Survey Questionnaire

The Reactions Survey used in the present study was intended to measure a variety of cognitive and affective reactions of evaluators toward the rating process. The survey is based in part upon the questions asked by Friedman and Cornelius (1976) to elicit reactions to the use of different rating scale formats. A copy of the Reactions Survey instrument used in this study can be found in Appendix G. The questionnaire was developed with the intention of measuring several potential classes of rater reactions to the evaluation process including: satisfaction, resistance, motivation to be accurate, and perceptions of actual accuracy. The survey consists of a manipulation check portion (items 1-6, page 2) followed by four sections distinguishable by differences in item
format and content, as follows: Section 1, multiple choice items (questions 7-10, pp. 2-3), Section 2, semantic differential items (questions 1-26, pp. 3-4), Section 3, multiple choice items (questions 1-3, p. 4), and Section 4, Likert-style items (questions 1-8, p. 5).

Eleven major constructs or categories of rater reactions were chosen a-priori and each was measured via one or more questionnaire items. Figure 1 lists these constructs and the corresponding survey items. Each of the constructs included in Figure 1 was viewed as an important component of the more general "rater reactions" concept. Each of the 11 constructs (except #10) was measured by at least two items with different question formats.

Measures of Rating Quality:

Numerous types of measures of rating quality have been used by past researchers. These were divided into two general groups of measures to be referred to here as "visible" ones, and "convergent validity" ones. The "visible" set of indexes can be further divided into a "distributional assumption" group and a "feedback value" group. Those measures that will be called distributional assumption ones are more often referred to simply as measures of rating "errors" (though this may be a misnomer) and include indexes of halo, leniency and variability. These have been given the "distributional assumption"
<table>
<thead>
<tr>
<th>Construct</th>
<th>&quot;Reactions Survey&quot; Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Motivation to rate accurately</td>
<td>Qs 1,5</td>
</tr>
<tr>
<td>2) Satisfaction w/peer evaluation system</td>
<td>Qs 3,11</td>
</tr>
<tr>
<td>3) Resistance to peer evaluation system</td>
<td>Qs 4,15</td>
</tr>
<tr>
<td>4) Tension-anxiety</td>
<td>Qs 9,16,18</td>
</tr>
<tr>
<td>5) General affective reactions</td>
<td>Qs 6,8,10,11,13,19,20,23</td>
</tr>
<tr>
<td>6) Capability to rate accurately</td>
<td>Qs 2,37</td>
</tr>
<tr>
<td>7) Difficulty in evaluating peers</td>
<td>Qs 26,40</td>
</tr>
<tr>
<td>8) Perceived accuracy of peer ratings</td>
<td>Qs 17,21,23,35</td>
</tr>
<tr>
<td>9) Feedback value of peer ratings</td>
<td>Qs 32,41</td>
</tr>
<tr>
<td>10) Usefulness for administrative decisions</td>
<td>Q 33</td>
</tr>
<tr>
<td>11) Future use of system favored</td>
<td>Q 39</td>
</tr>
</tbody>
</table>

FIGURE 1
CONSTRUCTS AND CORRESPONDING ITEMS FROM
THE REACTIONS SURVEY
label here because the extent to which they truly reflect rating accuracy rests upon several important assumptions about the true distribution of ratee ability. It is assumed for instance, that ratee ability is distributed normally and that rating scale dimensions are independent (see Warmke, 1977 for a review). Furthermore, it is assumed that raters whose ratings more closely conform to these distributional assumptions have rated more accurately. Neither assumption is necessarily true. Within some settings the distributional assumptions themselves are false (see Cascio, 1978, p. 321), and even when the assumptions themselves are true it is possible for raters to purposely conform to the distributional assumptions without rating more accurately (Borman, 1975).

Within the present study four measures which fit the visible "distributional assumption" mold will include: 1) halo A (the level of trait covariance within a rater's evaluations; Warmke, 1977), 2) halo B (the pooled variance of ratings across dimensions; Borman, 1975), 3) leniency (the mean of ratings given by each rater; Borman & Vallon, 1974), and 4) variability (the standard deviations within dimensions across ratees; Borman & Vallon, 1974).

Another group of "visible" measures of rating quality will be labeled "feedback value" measures. Interestingly, the feedback value of ratings has not been examined by past researchers as an index of rating
quality. However, two such measures appear useful and were used here. These included: 1) the number of behavioral descriptions used by a rater to justify or document the ratings given on a particular dimension, and 2) the number of suggestions offered to the ratee of ways the ratee might improve. The feedback value measures rest upon the logical assumption that ratings which are more thoroughly documented by examples of past ratee behavior, or which include more suggestions for how ratees can improve, are of higher quality. This assumption would probably only be violated in the event that a rater were to fabricate examples of behavior or list meaningless suggestions. The feedback value measures are highly visible, and it seems probable that a rater could purposely alter the quantity of documentation and number of suggestions given. It may be recalled that Hypothesis 1 predicted that "visible" measures of rating quality (of both the distributional assumption and feedback value varieties) would be successfully altered by accountability procedures.

Another group of rating quality measures will be labeled "convergent validity" indexes. These will include 1) accuracy (the convergence of ratings with objective measures of performance; Borman & Rosse; Note 1), and 2) interrater agreement (the average correlation of ratings by different raters of the same ratees; Warmke, 1977).
Borman and Rosse (Note 1) have suggested that the "accuracy" of ratings can be determined by their correspondence to objective measures of the same traits. However, experimenters seldom collect such objective data. This is unfortunate because the level of convergence of ratings with good objective measures is perhaps the best available indication of rating quality within either field or laboratory sites. Within the present study objective measures of both in-class test performance and physical proficiency were available objective measures having a very high conceptual overlap with the "Mastery of Subject Matter" and "Physical Fitness" rating dimensions. Thus, it was possible to compute the accuracy of ratings within the present experiment.

Interrater agreement is a second possible measure of convergent validity. It is generally agreed that convergence among alternate rating sources is a likely reflection of accuracy or truth. For example, Campbell and Fiske's (1959) multitrait multimethod matrix mode of measuring convergent validity is based in part upon the "truth in convergence" assumption. In addition, Borman, Hough and Dunnette (Note 7) have provided strong evidence that interrater agreement is related to accuracy of ratings (when accuracy is measured within controlled lab settings, and with available true scores). These authors note that interrater agreement can be improved as an index
of accuracy by increasing the number of different raters (up to 4) evaluating the same ratee. Finally, James (1974) has indicated that the construct validity of criteria is enhanced by a demonstration of convergence among several different measurement sources.

Despite the evidence suggesting the soundness of interrater agreement as an index of rating quality, there is some reason to be cautious in such an interpretation. It is well known that, however unlikely, it is possible for several individuals to agree on something and all be wrong (see Hakel, 1966). Therefore, though it may be reasonable to say that without agreement inaccuracy is assured, the presence of agreement does not insure accuracy.

Despite possible criticisms of both "accuracy" and interrater agreement measures of rating quality, these are seen as superior to the distributional assumption measures described earlier. Hypothesis 2 predicted that accountability procedures would fail to favorably affect either of these convergent validity measures.

In summary, the following six "visible" measures of rating quality will be used: 1) trait covariance halo, 2) pooled variance halo, 3) leniency, 4) variability, 5) number of behavioral descriptions, and 6) number of behavioral suggestions. In addition, two "convergent validity" measures of rating quality will include:
1) accuracy, and 2) interrater agreement.

Statistical Analyses

Hypothesis 1 of the present study predicted that procedures intended to alter rater perceptions of accountability would differentially influence the visible measures of rating quality. The most appropriate test consisted of a MANOVA with Review (R), Knowledge of Standards (KS), and Review and Knowledge of Standards interaction (R x KS) as the independent variables, and the six measures of rating quality as the dependent variables.

Hypothesis 2 predicted that the experimental variables (R, KS and R x KS) would not differentially affect "convergent validity" measures of rating quality such as measures of "accuracy" and interrater agreement. This hypothesis was tested via MANOVA with R, KS, and R x KS as independent variables and "accuracy" scores and interrater agreement indexes as the dependent variables.

Hypothesis 3 predicted that the experimental manipulations would differentially affect "rater-reaction" questionnaire responses. This hypothesis was tested once again with R, KS, and R x KS as the independent variables and with the eleven constructs representing different components of rater reactions as the dependent variables within a single MANOVA analysis.
In cases where the MANOVAs suggested for answering Hypothesis 1-3 were significant, univariate ANOVAs were explored for explaining these effects. Additional group comparisons were also sometimes warranted at this stage. Additional analyses investigated the levels of acquaintance of participants with the recruits they evaluated, and the relation between rater reactions and levels on the 8 major indexes of rating quality.
RESULTS

The results portion of the present study will include the following sections: 1) manipulation checks, 2) effects of experimental manipulations on halo, variability, leniency, and number of behavioral examples and suggestions, 3) effects of experimental conditions upon interrater agreement and accuracy, 4) effects of experimental procedures on rater reactions to the rating process, 5) relations among measures of rating quality and rater reactions to the evaluation process, 6) acquaintance survey, and 7) additional analyses.

MANIPULATION CHECKS

Most recruits perceived the experimental manipulations as had been intended. The first six questions of the Reactions Survey (Appendix G) served as manipulation check items. Table 1 lists the means and standard deviations for these questions. Questions 1 and 3 were expected to yield uniform responses of 1 across experimental groups because participants in all groups were told that a) their CE would see their ratings (question 1) and b) they would receive feedback about the ratings they received (question 3). Questions 2, 4, 5 and 6
<table>
<thead>
<tr>
<th>Question</th>
<th>Group 1 (Review + Standards)</th>
<th>Group 2 (Review)</th>
<th>Group 3 (Standards)</th>
<th>Group 4 (Control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) CE will see ratings = 1, will not see = 2</td>
<td>Mean 1.14, S.D. .36</td>
<td>Mean 1.08, S.D. .28</td>
<td>Mean 1.15, S.D. .38</td>
<td>Mean 1.08, S.D. .28</td>
</tr>
<tr>
<td>2) Own name not on ratings = 1, name on rating form = 2</td>
<td>Mean 1.79, S.D. .43</td>
<td>Mean 1.85, S.D. .38</td>
<td>Mean 1.00, S.D. .00</td>
<td>Mean 1.00, S.D. .00</td>
</tr>
<tr>
<td>3) Will receive feedback on ratings received = 1, will not = 2</td>
<td>Mean 1.14, S.D. .36</td>
<td>Mean 1.00, S.D. .00</td>
<td>Mean 1.00, S.D. .00</td>
<td>Mean 1.00, S.D. .00</td>
</tr>
<tr>
<td>4) Ratings made will be evaluated by CE = 1, will not = 2</td>
<td>Mean 1.06, S.D. .27</td>
<td>Mean 1.25, S.D. .45</td>
<td>Mean 1.92, S.D. .28</td>
<td>Mean 1.91, S.D. .30</td>
</tr>
<tr>
<td>5) CE expects high quality &amp; specified clear standards = 1</td>
<td>Mean 1.71, S.D. .91</td>
<td>Mean 2.31, S.D. .75</td>
<td>Mean 2.00, S.D. .91</td>
<td>Mean 2.58, S.D. .79</td>
</tr>
<tr>
<td>CE expects high quality &amp; specified no clear standards = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CE has not indicated clear expectations on standards = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) Rating quality will affect CE view: 1 = greatly, 2 = moderately,</td>
<td>Mean 2.29, S.D. .83</td>
<td>Mean 2.46, S.D. .88</td>
<td>Mean 3.46, S.D. .77</td>
<td>Mean 3.82, S.D. .40</td>
</tr>
<tr>
<td>3 = a little, 4 = none</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 2
F. VALUES & SIGNIFICANCE LEVELS OF EFFECTS OF REVIEW AND KNOWN STANDARDS CONDITIONS ON MANIPULATION CHECK ITEMS

<table>
<thead>
<tr>
<th>Effects of</th>
<th>On: Question (Q)1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review (R)</td>
<td></td>
<td>.16</td>
<td>.38 **</td>
<td>1.62</td>
<td>84.85 **</td>
<td>1.16</td>
</tr>
<tr>
<td>Knowledge of Standards (KS)</td>
<td>1.43</td>
<td>.78</td>
<td>1.62</td>
<td>2.16</td>
<td>4.33 *</td>
<td>1.55</td>
</tr>
<tr>
<td>R x KS</td>
<td></td>
<td>.28</td>
<td>.78</td>
<td>1.62</td>
<td>.34</td>
<td>.01</td>
</tr>
</tbody>
</table>

For all items df = 3, 45.
* = p < .05
** = p < .001

MS errors for Qs 1-6 respectively: 1).086, 2).089, 3).083, 4).091, 5).75, 6).59
were expected to result in significant differences in mean responses across groups due to differential exposure to review and known standards experimental conditions.

Table 2 demonstrates that responses to items 2, 4, 5 and 6 were significantly different across experimental groups. The significant review effect upon question 2 \( (F_{[3,45]} = 88.38, p \leq .001) \) indicated that recruits in all four groups correctly perceived whether or not their names were on the rating forms. On question 4 recruits correctly perceived that their ratings would be evaluated by their CEs \( (F_{[3,45]} = 84.85, p \leq .001) \), and on question 6 they correctly understood that the quality of their ratings would influence their CEs view of them \( (F_{[3,45]} = 31.16, p \leq .001) \).

The significant knowledge of standards (KS) effect on question 5 demonstrated that recruits correctly perceived their CEs standards of rating quality as clearer within the KS groups \( (F_{[3,45]} = 4.33, p \leq .05) \). Thus all manipulation check results were significant and in the anticipated direction. The review manipulation was also apparently stronger than the knowledge of standards manipulation.

**EFFECTS OF EXPERIMENTAL PROCEDURES ON HALO, VARIABILITY, LENIENCY & WRITTEN COMMENTS MEASURES OF RATING QUALITY**

Hypothesis 1 suggested that compliance with rating quality standards would increase as the number of
accountability components (review & knowledge of standards) increased. Rating quality standards were presented in such a way as to be measurable by conventional psychometric indicators of halo, variability, leniency and numbers of written comments made by raters. Definitions and methods of measuring these psychometric indicators were presented in the Methods portion of the present report (p. 52). The means and standard deviations for each of the experimental groups are presented in Table 3.

Hypothesis 1 was tested using a 2 x 2 (review x knowledge of standards) MANOVA with six "visible" dependent variables including variability, leniency, pooled variance halo, interdimension covariance halo, number of written descriptive comments and number of written suggestions. Table 4 presents the results of this 2 x 2 MANOVA and thus shows effects due to Review (R), Knowledge of Standards (KS) and R x KS interaction upon the six dependent variables.

The multivariate F-value for the Review effect is shown in the left hand portion of Table 4 and was not significant. Likewise review did not have significant univariate effects upon any of the six dependent variables.

The Knowledge of Standards (KS) manipulation did yield a marginally significant multivariate effect upon the combined dependent variables ($F_{6,44} = 2.11, p <
<table>
<thead>
<tr>
<th>Measures</th>
<th>Group 1 (Review &amp; Known Standards) (N=14)</th>
<th>Group 2 (Review) (N=13)</th>
<th>Group 3 (Known Standards) (N=13)</th>
<th>Group 4 (Control) (N=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leniency (Means)</td>
<td>Mean 6.36</td>
<td>5.85</td>
<td>5.87</td>
<td>6.61</td>
</tr>
<tr>
<td></td>
<td>S.D. .63</td>
<td>.57</td>
<td>.59</td>
<td>.67</td>
</tr>
<tr>
<td>Variability (S.D. S)</td>
<td>Mean 2.55</td>
<td>3.25</td>
<td>3.15</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.08</td>
<td>1.43</td>
<td>1.07</td>
<td>1.76</td>
</tr>
<tr>
<td>Halo (Pooled Variance)</td>
<td>Mean 1.82</td>
<td>1.72</td>
<td>2.02</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>S.D. .71</td>
<td>.76</td>
<td>.83</td>
<td>1.04</td>
</tr>
<tr>
<td>M.S. Interaction (Interdimension Covariance Halo)</td>
<td>Mean 1.62</td>
<td>1.61</td>
<td>1.85</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>S.D. .77</td>
<td>.68</td>
<td>.73</td>
<td>.97</td>
</tr>
<tr>
<td># Written Descriptors</td>
<td>Mean 61.14</td>
<td>43.92</td>
<td>59.30</td>
<td>45.31</td>
</tr>
<tr>
<td></td>
<td>S.D. 29.77</td>
<td>26.77</td>
<td>24.30</td>
<td>24.43</td>
</tr>
<tr>
<td># Written Suggestions</td>
<td>Mean 16.57</td>
<td>9.00</td>
<td>14.92</td>
<td>9.54</td>
</tr>
<tr>
<td></td>
<td>S.D. 19.03</td>
<td>8.10</td>
<td>13.73</td>
<td>12.86</td>
</tr>
</tbody>
</table>
### TABLE 4

**Multivariate and Univariate ANOVA F-Values for Review, Knowledge of Standards and Interaction Effects on Six Measures of Rating Quality**

<table>
<thead>
<tr>
<th>Sources of Variance</th>
<th align="right">(df= 6,44) Multivariate $F^s$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th align="right">(df=3,49) Univariate $F^s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review (R)</td>
<td align="right">.51</td>
<td>.09</td>
<td>.00</td>
<td>.72</td>
<td>.62</td>
<td>.00</td>
<td align="right">.02</td>
</tr>
<tr>
<td>Knowledge of Standards (K.S.)</td>
<td align="right">2.11*</td>
<td>.95</td>
<td>1.11</td>
<td>1.20</td>
<td>.49</td>
<td>4.60**</td>
<td align="right">2.79*</td>
</tr>
<tr>
<td>RxKS Interaction</td>
<td align="right">2.31**</td>
<td>.31</td>
<td>.30</td>
<td>.60</td>
<td>13.54****</td>
<td>.05</td>
<td align="right">.08</td>
</tr>
</tbody>
</table>

* $F^s$ is Pillai trace (V) appropriations to F-Values.  
Halo A = Pooled Variance Halo  
Halo B = Interdimension Covariance Halo  
Var. = Variability  
Len = Leniency  
# Descs. = Number of Descriptive Comments  
# Suggs. = Number of Suggestions  

---

**Significance Levels:**  
**** = $p < .001$  
*** = $p < .01$  
** = $p < .05$  
* = $p < .10$  
0.001
Exploration of the univariate effects indicated that KS significantly increased the number of descriptive comments used by raters to document ratings ($F[3,49] = 4.60, p \leq .037$), and also marginally affected the number of written suggestions by raters ($F[3,49] = 2.79, p \leq .101$).

The R x KS interaction produced the strongest multivariate effect upon the set of six dependent variables ($F[6,44] = 2.31, p \leq .0499$). This effect was traceable to a strong univariate effect upon leniency ($F[3,49] = 13.54, p \leq .0006$). The means in Table 3 indicate that the lowest mean ratings were given by subjects in the review only and known standards only groups (Groups 2, 3). Both the control (Group 4) and combined review and known standards group (Group 1) had higher levels of mean ratings (leniency). Potential explanations for these results are offered in the Discussion chapter.

Significant differences across groups were also found within a dependent variable not included in Table 3 (the number of words used to describe a recruit's present behavior). In addition to the number of phrases describing a recruit's behavior, the number of words written on the rating forms was also computed. The mean number of (descriptive) words written on the forms for groups 1, 2, 3 and 4 respectively was: 433, 275, 333, and 282. A univariate analysis using experimental groups
as the independent variable yielded a significant effect upon number of written words (F[3, 49] = 3.34, p = .0265). As can be seen by the means the high accountability group (Group 1) wrote considerably more on the rating scales than did the other groups.

Inter-judge agreement on the number of comments and words on a given rating scale was computed on a sample of 10% (5) of the rating forms. For the number of comments and suggestions written on the forms reliability was nearly perfect (r = .998, 1.00).

**Effects of Experimental Procedures Upon Interrater Agreement and Accuracy**

Hypothesis 2 suggested that accountability manipulations could influence certain classes of dependent variables (e.g. visible measures such as halo, leniency, and number of written comments) independent of effects upon other important measures of rating quality (i.e. interrater agreement and accuracy).

**Interrater Agreement.** Levels of interrater agreement were computed by correlating the ratings by different raters of the same ratees separately for each of the four dimensions on the rating scale (on physical fitness, compatibility, mastery of subject matter, and leadership). Table 5 presents the means and standard deviations for interrater agreement (IRA) by experimental groups.
### TABLE 5
MEANS AND STANDARD DEVIATIONS OF INTERRATER AGREEMENT MEASURES OF RATING QUALITY

<table>
<thead>
<tr>
<th>Measures</th>
<th>Group 1 (Review &amp; Known Standards) (N=6 pairs)</th>
<th>Group 2 (Review) (N=6 pairs)</th>
<th>Group 3 (Known Standards) (N=6 pairs)</th>
<th>Group 4 (Control) (N=5 pairs)</th>
<th>All Groups Combined (N=23 pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRA1 (Interrater Agreement on &quot;Physical Fitness&quot; dimension)</td>
<td>Mean 0.59</td>
<td>0.67</td>
<td>0.76</td>
<td>0.55</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>S.D. 0.25</td>
<td>0.34</td>
<td>0.24</td>
<td>0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>IRA2 (Interrater Agreement on &quot;Compatibility&quot; dimension)</td>
<td>Mean 0.51</td>
<td>0.65</td>
<td>0.40</td>
<td>0.18</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>S.D. 0.33</td>
<td>0.25</td>
<td>0.22</td>
<td>0.17</td>
<td>0.29</td>
</tr>
<tr>
<td>IRA3 (Interrater Agreement on &quot;Mastery of Subject Matter&quot; dimension)</td>
<td>Mean 0.51</td>
<td>0.55</td>
<td>0.55</td>
<td>0.47</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>S.D. 0.22</td>
<td>0.29</td>
<td>0.23</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>IRA4 (Interrater Agreement on &quot;Leadership&quot; dimension)</td>
<td>Mean 0.27</td>
<td>0.60</td>
<td>0.58</td>
<td>0.32</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>S.D. 0.28</td>
<td>0.25</td>
<td>0.16</td>
<td>0.28</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Table 6 presents the multivariate and univariate effects of Review (R), Knowledge of Standards (KS) and R x KS interaction upon the four measures of interrater agreement.

It can be seen from Table 6 that none of the multivariate effects influenced overall interrater agreement at conventional levels of significance. Because the multivariate Review effect failed to approximate a $p < .05$ level of significance, the reader should be cautious in interpreting the significant effect of review upon interrater agreement on the dimension of compatibility ($F[3,19] = 7.75, p \leq .01$).

However, the multivariate R x KS interaction effect ($F[4,16] = 2.66, p \leq .07$) is sufficiently strong to merit further exploration. This effect appears to be primarily due to a strong R x KS interaction effect upon interrater agreement on the "Leadership" dimension ($F[3,19] = 7.99, p \leq .01$) The means (Table 5) on this dependent variable suggest that both the Review only and Known Standards only groups (Groups 2,3) were in greater agreement than either the control (Group 4) or review and known standards group (Group 1).

Table 6 also indicates a marginally significant univariate effect of R x KS interaction upon agreement on the "compatibility" dimension ($F[3,19] = 3.04, p \leq .097$). In this case the means suggest that the review only
### TABLE 6
MULTIVARIATE AND UNIVARIATE ANOVA F-VALUES FOR REVIEW, KNOWLEDGE OF STANDARDS AND R x KS INTERACTION EFFECTS UPON FOUR MEASURES OF INTERRATER AGREEMENT

<table>
<thead>
<tr>
<th>Sources of Variance</th>
<th>Multivariate Fs&lt;sup&gt;a&lt;/sup&gt; (df=4,16)</th>
<th>IRA1 (Fitness)</th>
<th>IRA2 (Compatibility)</th>
<th>IRA3 (Classroom Grades)</th>
<th>IRA4 (Leadership)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review (R)</td>
<td>1.87&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.05</td>
<td>7.75***</td>
<td>.04</td>
<td>.02</td>
</tr>
<tr>
<td>Knowledge of Standards (KS)</td>
<td>.95</td>
<td>.37</td>
<td>.14</td>
<td>.03</td>
<td>.10</td>
</tr>
<tr>
<td>R x KS Interaction</td>
<td>2.66*</td>
<td>1.88</td>
<td>3.04*</td>
<td>.34</td>
<td>7.99***</td>
</tr>
</tbody>
</table>

<sup>a</sup>Pillai Trace (V) approximations to F-values

IRA = Interrater Agreement

*** = p < .01
** = p < .05
* = p < .10
b = p < .16
condition (Group 2) is most favorable to agreement while the control condition (Group 4) is least favorable.

In summary, the effects of the experimental manipulations upon interrater agreement were not straightforward, nor were they strong. However, an $R \times KS$ interaction effect did appear to be present, especially within agreement on the "leadership" dimension.

**Accuracy.** Accuracy was computed by correlating the ratings given to several ratees with the objective scores earned by the same ratee on the corresponding rating dimension. In the present case two indexes of accuracy were computable. The first involved correlating scores on the "Physical Fitness" dimension with scores received by the same recruits on a physical fitness scale which consisted of a composite of six objective measurements (time in 1½ mile run, pushups, situps, resting heart rate, grip strength, and agility run). The second accuracy measure involved correlating scores on the "Mastery of Subject Matter" dimension with the mean scores earned by recruits across twelve major in-class written examinations.

Table 7 presents the means on Accuracy 1 (Physical Fitness) and Accuracy 2 (Mastery of Subject Matter). Table 8 shows the results of a MANOVA analysis giving the effects of $R$, $KS$ and $R \times KS$ interaction on both Accuracy 1 and Accuracy 2.
<table>
<thead>
<tr>
<th>Measures</th>
<th>Group 1 (Review &amp; Known Standards) (N=14)</th>
<th>Group 2 (Review) (N=13)</th>
<th>Group 3 (Known Standards) (N=13)</th>
<th>Group 4 (Control) (N=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy 1 (Physical Fitness)</td>
<td>Mean .65</td>
<td>.69</td>
<td>.58</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>S.D. .16</td>
<td>.22</td>
<td>.16</td>
<td>.24</td>
</tr>
<tr>
<td>Accuracy 2 (Mastery of Subject Matter)</td>
<td>Mean .41</td>
<td>.49</td>
<td>.59</td>
<td>.44</td>
</tr>
<tr>
<td></td>
<td>S.D. .51</td>
<td>.19</td>
<td>.24</td>
<td>.34</td>
</tr>
<tr>
<td>Sources of Variance</td>
<td>Multivariate Fs (df= 2,45)</td>
<td>Univariate Fs (df=3,46)</td>
<td>Accuracy 1 (Physical Fitness)</td>
<td>Accuracy 2 (Mastery of Subject Matter)</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-----------------------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Review (R)</td>
<td>2.77*</td>
<td>5.26**</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>Knowledge of Standards (KS)</td>
<td>.07</td>
<td>.05</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>R x KS Interaction</td>
<td>1.09</td>
<td>1.00</td>
<td>1.21</td>
<td></td>
</tr>
</tbody>
</table>

*** = p < .01
** = p < .05
* = p < .10
As can be seen from Table 8 only one multivariate effect upon overall accuracy demonstrated effects at a marginally significant level. Review did appear to influence overall accuracy ($F(2,45) = 2.77, p \leq .07$). As can be seen in Table 8 this effect was due entirely to a univariate effect of review upon Accuracy 1 (Physical Fitness) ($F[3,46] = 5.26, p \leq .027$). The means in Table 7 suggest that Review did tend to increase Accuracy 1.

In summary, Accuracy appeared to be somewhat increased by the review component of accountability. However, this was true only for accuracy on the "Physical Fitness" dimension and not for the "Mastery of Subject Matter" dimension. Interestingly, the correlation between Accuracy 1 and Accuracy 2 measures was very low ($r = .023$).
Effects of Experimental Procedures on Rater Reactions to the Rating Process

Rater reactions to the evaluation process were measured by a "Reactions Survey" described earlier (see Appendix G). Analysis of questionnaire responses involved several steps. First, items which were related to potential rater reaction content domains were isolated and combined. The eleven a-priori item groupings which resulted were presented earlier (Figure 1, page 53). The use of a-priori item groupings (constructs) was chosen over factor analytic procedures because of the large number of questionnaire items (N=42) relative to participants (N = 53).

Second, item responses were standardized within experimental groups so that items with more response categories (e.g. Q.5 with 7 response categories) would be weighed the same as those with fewer (e.g. Q1 with 5 response categories). Standardizing all item means and variance insured that all items would be weighed equally when they were combined into construct groupings.

Third, the correlations among all individual standardized items and among the 11 standardized rater reaction constructs were computed, and the internal consistency reliability of the 11 constructs was ascertained using Cronback's coefficient alpha computational formula. The internal consistency reliabilities
of the 11 constructs were moderately low (ranging from .53 to .90) for nine of the 11 constructs, and internal reliabilities were not computable for two constructs because they were measured using single items. Furthermore, the correlations among the constructs themselves were generally quite high, suggesting that some of the constructs were measuring the same thing. For example, the unadjusted correlation between the initial constructs (see p. 53) called "satisfaction with peer evaluation system" and the one called "general affective reactions" was +.82. When the "true" correlation between these two constructs was estimated (using the Spearman-Brown Prophecy Formula), the resulting correlation between the constructs was .96 suggesting that the two constructs should be combined. Further analyses utilizing the Spearman-Brown Formula led to the combining of several of the initial constructs into a new set of relatively independent constructs. Figure 2 lists the resulting set of five revised constructs, their internal consistency reliabilities, and the items each construct was composed of. As can be seen from Figure 2, the internal consistencies of the revised set of constructs was generally high (.75-.93).

Table 9 lists the correlations among the revised set of five rater reaction constructs. As can be seen from Table 9 the correlations range from a low of .15
<table>
<thead>
<tr>
<th>Constructs</th>
<th>Internal Consistency Estimates*</th>
<th>&quot;Reactions Survey&quot; Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Motivation to rate accurately</td>
<td>.88</td>
<td>Q1, Q5, Q12, Q30</td>
</tr>
<tr>
<td>2) Dissatisfaction with peer rating system</td>
<td>.93</td>
<td>Q3, Q6, Q8, Q10, Q11, Q19, Q20, Q28, Q38</td>
</tr>
<tr>
<td>3) Perceived Accuracy of Peer Ratings</td>
<td>.88</td>
<td>Q17, Q21, Q23, Q24, Q27, Q31, Q35</td>
</tr>
<tr>
<td>4) Resistance to Peer Evaluation System</td>
<td>.79</td>
<td>Q4, Q15, Q36</td>
</tr>
<tr>
<td>5) Future Use of System Favored</td>
<td>.75</td>
<td>Q33, Q39</td>
</tr>
</tbody>
</table>

*Cronbach's Coefficient Alpha

FIGURE 2

REVISED CONSTRUCTS, INTERNAL CONSISTENCY ESTIMATES, AND CORRESPONDING ITEMS FROM THE "REACTIONS SURVEY"
<table>
<thead>
<tr>
<th></th>
<th>MOT</th>
<th>DISA</th>
<th>ACC</th>
<th>RESI</th>
<th>FUTU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOT</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISA</td>
<td>-0.57</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.48</td>
<td>-0.51</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESI</td>
<td>-0.49</td>
<td>0.66</td>
<td>-0.22</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>FUTU</td>
<td>0.36</td>
<td>-0.63</td>
<td>0.15</td>
<td>-0.49</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Note: Correlations of .35 or greater are significant at the p < .01 level. Correlations of .30 or greater are significant at the p < .05 level. MOT = Motivation, DISA = Dissatisfaction, ACC = Accuracy, RESI = Resistance, FUT = Future.
(between the "future use of system favored" construct and the "perceived accuracy of peer ratings" construct) to a high of .66 (between the "dissatisfaction with peer evaluation system" construct and the "resistance to peer evaluation system" construct). When the .66 correlation is corrected for unreliability using the Spearman-Brown formula, the resulting correlation is .77. Thus, though some of the five constructs do overlap considerably, they were viewed as sufficiently distinct to be reported and analyzed separately.

The reader interested in the raw means of "Reaction Survey" items by experimental group is referred to Appendix H. For purposes of statistical analysis the raw means and variances of items and the resulting constructs were standardized. This was done to insure that differences in item variances and in the number of items comprising each construct did not serve to distort construct means and standard deviations. Table 10 lists the standardized means and standard deviations of the five "Reactions Survey" constructs by experimental groups. Means were standardized at 5.0 and standard deviations at 1.0 across experimental groups. Inspection of the means in Table 10 suggested the possibility that differences by experimental group might be found for the "motivation" and "future" constructs.
<table>
<thead>
<tr>
<th>Construct</th>
<th>Group 1 (Review &amp; known Standards)</th>
<th>Group 2 (Review only)</th>
<th>Group 3 (Known standards only)</th>
<th>Group 4 (Control)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean 5.52 N=14</td>
<td>Mean 4.66 N=13</td>
<td>Mean 4.94 N=13</td>
<td>Mean 4.85 N=13</td>
</tr>
<tr>
<td></td>
<td>S.D. .50</td>
<td>S.D. 1.00</td>
<td>S.D. .90</td>
<td>S.D. 1.33</td>
</tr>
<tr>
<td>Motivation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disatisfaction</td>
<td>Mean 4.91 N=14</td>
<td>Mean 5.36 N=13</td>
<td>Mean 4.92 N=13</td>
<td>Mean 4.80 N=13</td>
</tr>
<tr>
<td></td>
<td>S.D. .99</td>
<td>S.D. 1.09</td>
<td>S.D. .80</td>
<td>S.D. 1.11</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Mean 5.17 N=14</td>
<td>Mean 4.99 N=13</td>
<td>Mean 4.81 N=13</td>
<td>Mean 5.02 N=13</td>
</tr>
<tr>
<td></td>
<td>S.D. .65</td>
<td>S.D. 1.17</td>
<td>S.D. .93</td>
<td>S.D. 1.25</td>
</tr>
<tr>
<td>Resistance</td>
<td>Mean 4.79 N=14</td>
<td>Mean 5.33 N=13</td>
<td>Mean 5.10 N=13</td>
<td>Mean 4.80 N=13</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.14</td>
<td>S.D. 1.20</td>
<td>S.D. .47</td>
<td>S.D. 1.02</td>
</tr>
<tr>
<td>Future</td>
<td>Mean 5.00 N=14</td>
<td>Mean 4.52 N=13</td>
<td>Mean 5.04 N=13</td>
<td>Mean 5.44 N=13</td>
</tr>
<tr>
<td></td>
<td>S.D. .98</td>
<td>S.D. 1.11</td>
<td>S.D. .80</td>
<td>S.D. .98</td>
</tr>
</tbody>
</table>

Note: Construct values were standardized to have means of 5.0 and S.D.s of 1.0.
To determine whether the mean differences in Table 10 were significant, multivariate and univariate tests were applied. These results are shown in Table 11. Within the present context Sarle (Note 7) has suggested that multivariate effects significant at the $p \leq .10$ level are sufficient for interpretation. This is based on the small sample size in the present study and the consequent low power of these MANOVAs. In addition, failure to interpret effects in the $p \leq .05 \rightarrow .10$ range would probably result in an unwarrantedly high risk of a Type II error. Even given this more liberal $p \leq .10$ significance level, none of the multivariate effects shown in Table 11 were sufficient to warrant interpretation. Sarle (Note 7) has also suggested that when multivariate effects do not reach interpretable significance levels, univariate effects should be interpreted only if they are strong enough to preserve the overall experimentive error rate at the $p \leq .10$ level. Therefore, in Table 11 the five univariate effects should be interpreted only if they reach a $p \leq .02$ ($.10$ divided by 5) level of significance. The BonFerroni post hoc tests described in Kirk (1968) apply this same logic of guarding against large levels of experimentive error rate. Given this restriction, none of the univariate effects in Table 11 was sufficiently strong to be interpreted.
<table>
<thead>
<tr>
<th>Sources of Variance</th>
<th>Multivariate Fs(^a) (df=5,45)</th>
<th>Univariate Fs (df=3,49)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MOT</td>
<td>DISA</td>
</tr>
<tr>
<td>Review (R)</td>
<td>1.25</td>
<td>1.04</td>
</tr>
<tr>
<td>Knowledge of Standards (KS)</td>
<td>.90</td>
<td>3.16*</td>
</tr>
<tr>
<td>R x KS Interaction</td>
<td>.86</td>
<td>2.06(^b)</td>
</tr>
</tbody>
</table>

\(^a\) Pillai trace (V) F Approximations

*** = \( p \leq .01 \)
**  = \( p \leq .05 \)
*   = \( p \leq .10 \)
\(^b\) = \( p \leq .15 \)
Though none of the effects shown in Table 11 were strong enough to be legitimately interpreted, the reader may be interested to note the directions of the means in Table 10 (and the almost interpretable effects in Table 11) were all in the hypothesized directions. The trends of the means suggested that Review did tend to lead to more negative rater reactions with respect to "Desire to use the system in the future," and Knowledge of Standards did seem to vaguely increase rater motivation to rate accurately. Nonetheless, the reader is cautioned that these were only weak trends which were not worthy of legitimate interpretations.

Overall, the effects of the experimental conditions upon the five rater reaction constructs were not strong. However, the trends of the means for the reaction's constructs were consistently yet weakly in the directions suggested by Hypothesis 3.

Relations Among Measures of Rating Quality and Rater Reactions to the Evaluation Process

The correlational analyses to be presented in this section touches upon two important questions. First, what are the correlations among the rating quality indexes used here? Second, what are the correlations between measures of rating quality (which reflect actual rater behavior) and measures of rater reactions (which reflect rater attitudes)?
The first question was answered by computing a correlational matrix for all measures of rating quality. This is shown in the upper portion of Table 12. These results show that only two correlations among the "visible" measures of rating quality were significant. Variability was highly correlated with interdimension covariance halo ($r = .46, p \leq .008$) and the number of comments written on the rating forms was highly correlated with the number of suggestions offered to the ratee ($r = .51, p \leq .0002$). The correlations between the two measures of accuracy and the "visible" measures of rating quality were generally quite low. Accuracy 1 (on the Physical Fitness dimension) was correlated near zero ($-.08 \rightarrow .10$) with the other measures, and Accuracy 2 (on the Mastery of Subject Matter dimension) was also not significantly correlated with any of the "visible" measures of rating quality. This analysis lends support to the theoretical rationale accompanying Hypothesis 3. It was suggested in this rationale that accuracy might be independent of the "visible" ("distributional assumption" and "feedback value") measures of rating quality. The present results support the independence of "visible" measures and accuracy measures.

The relationship between all rating quality measures (rater behavior) and rater subjective reactions (rater attitudes) can be found in the lower portion of
<table>
<thead>
<tr>
<th></th>
<th>Hal. 2</th>
<th>Var.</th>
<th>Len.</th>
<th>Commo.</th>
<th>Sugno.</th>
<th>Acc. 1</th>
<th>Acc. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAL 2</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>.46</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEN</td>
<td>.19</td>
<td>.05</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMNO</td>
<td>.09</td>
<td>.26</td>
<td>-.10</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUGNO</td>
<td>.26</td>
<td>.12</td>
<td>.12</td>
<td>.51</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC 1</td>
<td>-.03</td>
<td>-.08</td>
<td>-.04</td>
<td>.01</td>
<td>.10</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>ACC 2</td>
<td>-.10</td>
<td>.18</td>
<td>.19</td>
<td>.13</td>
<td>.01</td>
<td>-.01</td>
<td>1.0</td>
</tr>
<tr>
<td>MOT</td>
<td>.09</td>
<td>.33</td>
<td>-.13</td>
<td>.40</td>
<td>.14</td>
<td>-.11</td>
<td>.06</td>
</tr>
<tr>
<td>DISA</td>
<td>-.19</td>
<td>.44</td>
<td>-.13</td>
<td>-.31</td>
<td>-.15</td>
<td>.08</td>
<td>.03</td>
</tr>
<tr>
<td>ACC</td>
<td>.18</td>
<td>.33</td>
<td>.08</td>
<td>.28</td>
<td>.12</td>
<td>-.13</td>
<td>-.13</td>
</tr>
<tr>
<td>RES</td>
<td>-.02</td>
<td>-.26</td>
<td>-.12</td>
<td>-.20</td>
<td>-.09</td>
<td>-.08</td>
<td>-.03</td>
</tr>
<tr>
<td>FUTU</td>
<td>.09</td>
<td>.37</td>
<td>.09</td>
<td>.25</td>
<td>.22</td>
<td>-.07</td>
<td>-.06</td>
</tr>
</tbody>
</table>

Note: Correlations greater than .34 are significant at the p ≤ .01 level. Correlations greater than .27 are significant at the p ≤ .05 level. Hal 2 = Interdimension covariance halo, Var = Variability, Len = Leniency Commo = Number of written comments, Sugno = Number of written suggestions Acc 1 = Accuracy on physical fitness, Acc 2 = Accuracy on Mastery of Subject Matter, Mot = Motivation, Disa = Dissatisfaction, Acc = Accuracy, Res = Resistance, Futu = Future.
Table 12. These results show that the relationship between rater attitudes and behavior was generally low. However there were two definite exceptions to this and several of these results are quite interesting.

Of the seven measures of rating quality (rater behavior) used, only two (variability, and number of comments) were correlated at moderate levels with rater subjective reactions (rater attitudes). However, variability (the extent to which a rater clearly distinguished better performance from poorer performance) was correlated at least at the $p \leq .06$ level of significance with all five of the rater reaction constructs. Raters with attitudes favorable to high quality ratings (e.g. high motivation, low dissatisfaction, high perceptions of accuracy, low resistance, and favoring future use of peer ratings) tended to consistently have low levels of variability errors. In other words, raters with favorable attitudes were more willing to "spread out" ratees so that good performers were clearly distinguishable from poor ones.

Very similar results to those for the variability measure were obtained for the "number of comments" measure of rating quality. Raters with attitudes more favorable toward peer ratings tended to also write more comments on the rating forms to document their ratings. The consistency of the correlational results suggests
that some kinds of rater behavior (writing comments on
the form and "spreading out" ratees) are consistently
and strongly related to rater attitudes.

Several other interesting results can be found in
the lower portion of Table 12. First, leniency, which
has been called a largely motivational error by Decker
(1977) appears in the present case not to be highly
related to the rater motivation construct \(r = -.13\).
Likewise, leniency appears to be generally unrelated to
other rater attitude constructs \(rs\) range from -13 to
+.08). Second, the relation between both measures of
rater accuracy and the five attitude constructs was low
(all correlations were insignificant and ranged from
-.13 to +.08). Finally, one curious result was that a
rater's subjective perception that he/she was rating
accurately was negatively correlated with actual
accuracy \(rs = -.13\). Overall, these results suggest
that rater attitudes were consistently associated with
variability and number of comments measures of rating
quality but were unrelated to rating accuracy and/or
halo, number of suggestions, and leniency measures.

**Acquaintance Survey**

The "Acquaintance Survey" used in the present study
was administered to participants in order to determine
two things: First, in an absolute sense how well
acquainted were recruits with their associates and second, were recruits better acquainted with recruits in their own CE group (who served as ratees) than with recruits outside of their CE group?

The scales used to determine both absolute and relative levels of acquaintance were 7 point Likert-type items (see Appendix B). Each recruit evaluated how well he/she knew each other recruit in the class by circling a number on a scale ranging from 1 ("not at all," to 4 ("moderately well") and 7 ("extremely well"). The overall mean level of acquaintance, summed across all recruits, was 3.2 on the 7 point scale, suggesting that recruits knew one another slightly less than "moderately well" one month prior to making peer ratings.

A comparison of the levels of acquaintance within and outside of an individual's own CE group yielded a mean of 3.1 for recruits outside the group and 3.7 for those within the group. These findings suggest that the CE group was a meaningful administrative and social grouping in the academy and that recruits knew the recruits they were asked to evaluate somewhat better than the remainder of the class. Thus, the use of other members of a recruit's CE group as ratees resulted in slightly higher levels of acquaintance (and greater potential for high quality ratings) than would have resulted from the random assignment of ratees.
The levels of acquaintance with members of one's own 
CE group (ratees) was highly uniform across experimental 
groups (as would be expected from the random assignment 
of recruits to experimental groups). For groups 1, 2, 3 
and 4 the mean level of acquaintance were 3.98, (S.D. = 
1.1), 3.5 (S.D. = 1.1), 3.6 (S.D. = 1.1) and 3.9 (S.D. 
= 1.2) respectively. These results suggest that con-
 founding of results due to acquaintance levels with 
ratees was unlikely.

Additional Analyses

A limited number of post hoc multiple group 
comparisons were carried out which may be of interest to 
the reader. It is appropriate to carry out such post hoc 
group comparisons when a significant multivariate inter-
action effect is found. It may be recalled that in Table 
4 a significant Review by Known Standards effect was 
found. In this case the univariate interaction effect 
on leniency was sufficiently strong to also warrant 
interpretation. As noted in Kirk (1968, p. 69-98), Dunn's 
(also called Bonferroni t ) multiple group comparisons 
are appropriate, as are Sheffe's, post hoc tests. It is 
appropriate to report the results of whichever of these 
two tests is most powerful. In the case of the leniency 
interaction effect the Bonferroni values indicated that 
Group 2 (Review only) and Group 3 (Known Standards only) 
complied to a greater degree than did Group 4 (Control).
These results were significant at the $p \leq .05$ level. However, Group 1 (R + KS) was not significantly different from either Group 2 and 3 or from Group 4.

A second significant multivariate interaction effect can be found in Table 6. In this case the interaction effect of R x KS on interrater agreement on the "leadership" dimension bears interpretation. In this case Group 2 (Review only) and Group 3 (Known Standards only) had higher interrater agreement than both Group 4 (Control) and Group 1 (R + KS), at the $p \leq .05$ level.

No other significant interaction effects were found within the study which bear directly on the hypotheses. Nonetheless, some other post hoc tests were feasible and may help the reader to better understand some of the results which did not directly test hypotheses. For example, the right hand column of Table 5 suggests that the mean levels of interrater agreement may have been higher for the two "objective" dimensions (Physical Fitness, Mastery of Subject Matter) than the two "subjective" dimensions (Compatibility, Leadership). Post hoc tests showed that the combined objective dimensions were higher in agreement than the combined subjective dimensions. The value was significant at the $p \leq .05$ level of significance. In addition, interrater agreement on the Physical Fitness dimension was significantly higher than on either the Compatibility or the Leadership
dimensions ($p \leq .05$) but not higher than on the Mastery of Subject Matter dimension. The reader interested in the Fortran computer programs used to compute the dependent variables shown in Table 3 (halo, leniency variability), Table 5 (interrater agreement) and Table 7 (accuracy) can find these programs listed in Appendix I.
DISCUSSION

The research reported in the previous chapters focused upon the impact of variables related to the construct of accountability (Review and Knowledge of Standards) upon various psychometric measures of rating quality and upon perceptual measures of rater reactions to the evaluation task. The present chapter is divided into 4 sections. First, the findings with respect to the major hypotheses are reviewed; second, explanations for some of the unexpected results are proposed. Third, potential limitations of the present study are discussed. Fourth, conclusions and implications for future research are suggested.

1) Summary of Major Findings:

Hypothesis 1. Hypothesis 1 predicted that as the number of accountability components (Review and Knowledge of Standards) increased, compliance with rating quality standards would also increase. Compliance level was directly measurable through several "visible" indicators of rating quality including halo, leniency, variability, number of comments and number of suggestions. The hypothesis received partial support.
Both of the single component groups (Review only, and Known Standards only) did generally outperform the control group. However, the group containing both accountability components generally failed to outperform the control. The analysis which was performed on the present data was directed toward ascertaining whether the presence of either accountability component was influential in increasing compliance with rating quality standards. The results shown in Table 4 (p. 66) show that Knowledge of Standards did tend to increase compliance, at least with the standards suggesting increased written documentation on the rating form. (Both the number of written comments on the form and (marginally) the number of suggestions for ratees were increased by the Known Standards component of accountability.

Also related to Hypothesis 1 was the finding of an (R x KS) interaction effect upon leniency error in Table 4. Both single component groups outperformed the control group, but the dual component group was somewhere in between the control and the single component groups and was not significantly different from either.

In summary, the obtained results provided support for the efficacy of Knowledge of Standards alone in increasing compliance to rating quality standards which emphasized that ratings be documented with several descriptive comments and suggestions and that mean ratings
not be overly high. However, Review alone produced
greater compliance only to the leniency standard (pro-
ducing lower mean ratings for Group 2). Furthermore, the
combination of Knowledge of Standards and Review did not
increase levels of compliance and in fact reduced
compliance when compared to the single component groups.

These findings suggest that in the present case
Knowledge of Standards alone (essentially a training
condition) was most effective in increasing rater
compliance with easily visible standards of rating
quality.

Hypothesis 2. Hypothesis 2 predicted that account-
ability procedures (including Review and Knowledge of
Standards components) would not significantly increase
the convergent validity of ratings. Convergent validity
was measured by pooled interrater agreement (on the
four rating dimensions) within experimental groups and
by an accuracy index reflecting the convergence of ratings
with objective scores on the dimensions. This hypothesis
also implied that a study might reach different con-
cclusions about the effectiveness of experimental mani-
pulations depending on whether the "visible" or "con-
vergent validity" indexes of rating quality were used.
This hypothesis too was only partially supported.

The results shown in Tables 6 and 8 (pp. 71, 74)
indicate that the Review component of accountability
did serve to increase interrater agreement on the "compatibility" dimension and accuracy on the "Physical Fitness" dimension. In addition, Table 6 shows an R x KS interaction effect on interrater agreement on the "Leadership" dimension. The interaction in this case showed the same pattern as the interaction effect on leniency. Once again the single component groups outperformed the control and were also superior to the dual component group.

Taken together, the effects of accountability procedures upon convergent validity (interrater agreement and accuracy) appear to be moderated by the rating dimension being evaluated. However, the groups consisting of Review alone (Group 2) and Known Standards alone (Group 3) had higher levels of both interrater agreement and accuracy than the group with Review and Known Standards (Group 1) and the Control (Group 4).

These results suggest that the presence of accountability components is related to convergent validity, contrary to Hypothesis 2. However, the additional implication of Hypothesis 2, that the conclusions of a study might be altered depending on whether the "visible" or "convergent validity" measures were used as criteria, was largely supported. The reader may take note that based on the "visible" indexes alone he/she would conclude that the Known Standards manipulation alone is
most effective for increasing rating quality. However, based on the "convergent validity" measures he/she would conclude that the Review alone manipulation was most effective. This difference in the conclusions drawn from the study is of considerable practical importance because "visible" measures of rating quality almost totally dominate the performance appraisal field, and because convergent validity measures are generally considered to be of greater theoretical and practical value.

In conclusion, the portion of Hypothesis 2 which predicted no accountability effects upon convergent validity was not supported. However, the implication of Hypothesis 2 that "convergent validity" and "visible" measures of rating quality are affected differently by accountability procedures, was supported.

Hypothesis 3. Hypothesis 3 predicted that rater reactions to the rating process would be "more negative" with the review component present than without it. In addition, Hypothesis 3 predicted more negative reactions for the Review only group (Group 2) than for the Review with Known Standards group (Group 1). This hypothesis was largely not supported by the obtained results. Though the trends of the means shown in Table 10 (p. 81) were almost entirely in the expected directions, none of these results reached acceptable levels of significance.
In summary, the "feedback value" (number of written comments and suggestions) was improved by the Known Standards component of accountability. Furthermore, the Review component of accountability served to increase interrater agreement on one of the four dimensions ("compatibility") and accuracy on the "physical fitness" dimension. Finally, the accountability components failed to significantly influence rater subjective reactions to the evaluation process.

Unexpected Results and Potential Explanations

Perhaps the major unexpected finding within the present experiment was the general high quality of ratings across all experimental (and control) groups. This uniform high quality of ratings may also be responsible for the moderate to weak support provided for Hypothesis 1, which predicted differences in "visible" levels of rating quality across groups.

Prior to explaining why the quality of ratings was uniformly high, some support should be provided for the contention itself. The ratings are believed to have been of general high quality because: First, the levels of accuracy across groups averaged .55 and the levels of interrater agreement averaged slightly above .5. Both of these figures suggest that the ratings given by recruit peers within the present context captured a great
deal of the true variance in performance. Second, the means on the "visible" measures of rating quality were quite impressive when compared to at least one other published report which used comparable measures. When the mean levels for both measures of halo and variability in the present study are compared to the levels in the Warrake & Billings (1979) study, the superiority of overall rating quality here is apparent. In fact, on these three rating quality indexes even the lowest quality ratings in the present experiment (by the control group) were of superior quality to the highest quality ratings (by the scale construction group) in the previous Warrake & Billings (1979) study. The comparison of the levels of these visible measures of rating quality to those in the previous study should add credibility to the claim that ratings by all groups of raters within this study were of high quality.

Several reasons for the high quality ratings by all of the present experimental groups appear reasonable. Most of the reasons can be viewed as contextual in nature. Contextual factors likely influencing the high quality ratings included 1) the para-military training setting which encouraged conformity to the expectations of supervisors, 2) the clearly stated expectations of accurate ratings conveyed by both the Training Bureau Commander and CEs "R" and "M", 3) the fact that most
recruits had no prior experience with peer ratings and were thus being exposed to a novel stimulus, the fact recruits were told that the ratings would be seen by their supervisors and could influence their own potential for promotion. All of these factors combined to create a context or climate which was very conducive to high levels of rater motivation to be conscientious and accurate. Some of these contextual factors are discussed at more length in an upcoming section dealing with the generalizability of the obtained results.

It should be emphasized that all of the above-mentioned contextual factors influenced all experimental groups within the present study, including the control group. Thus, far from having a control group which was unmotivated to rate accurately, quite the opposite appears to have been the case.

The contention that rater motivation was probably high for all rater groups is supported by the high levels of self-reported motivation to rate accurately on the "Reactions Survey." For example, 100% of the raters in group 1 claimed to be highly motivated to be accurate but most control group raters also claimed to be either "highly" or "somewhat" motivated. Thus, though motivational increases may have been created by the accountability manipulations, the levels of motivation created by the overall context of the ratings was so high to
begin with that the effects of accountability were bounded by these ceiling effects.

Given the potential strong impact of the contextual factors noted above on rater motivation and rating accuracy, perhaps it is not surprising that the accountability manipulations which were used produced relatively minor effects.

This would suggest that in order for accountability manipulations to have a maximal impact upon rater motivation and accuracy, contextual variables should be such that relatively low initial levels of motivation to rate accurately exist. Only when ratings by a control group can be expected to yield relatively low quality ratings would the implementation of accountability manipulations designed to increase rater motivation be expected to have a large impact.

The presence of important contextual variables in the present setting appear to satisfactorily explain the uniformly high quality ratings by all experimental groups and the absence of strong effects on "visible" measures of rating quality by the accountability manipulations used. However, a second unexpected result within the present study is more difficult to explain.

The negative outcomes obtained for the Review with Known Standards group (when compared to the Review only and Known Standards only groups) is difficult to explain.
One possible explanation may lie in a number of operational problems with the combined Review and Known Standards group, which made it different than a simple combination of the two single component groups. Differences between the two component and single component groups included the following: 1) The two component group was administered from 1-3 P.M. while the single component groups were administered from 3-5 P.M., 2) Due to inexperience by the CE, the instructions given to the two component group were presented in a choppy and non-natural manner, 3) Also due to inexperience by the CE, the instructions given to the two component group were incomplete (e.g. the Review instructions portion of the presentation was, inadvertently, almost totally overlooked and the Knowledge of Standards instructions were quite terse), and 4) the present author was, inadvertently, much more visible to participants in the single component groups because he appeared during the time instructions were being presented to turn on the tape recorders for these two experimental sessions. These differences between the single component groups and the dual component group may partially explain the ineffectiveness of the dual component group.

Two other possible explanations for the failure of the dual component group are also possible. First, it is possible that the theoretical formulation of
accountability presented in Chapter 1 is erroneous. Perhaps accountability is something more (or less) than the proposed components of Knowledge of Standards and Review. This explanation seems unlikely, however, because the self reported levels of motivation to rate accurately were at the highest level for the dual component group, consistent with theoretical expectations. Thus, the self-reported motivation levels would suggest that raters in the dual component group failed to comply with the rating quality standards because these standards were not clear to them.

As was noted earlier the Knowledge of Standards instructions presented to the dual component group were quite terse. It is possible that these were not explained fully enough to clarify the standards that would be used to evaluate the ratings.

A final possible explanation for the lower compliance to rating quality standards by the dual component group is that their very high levels of self reported motivation to rate accurately were accompanied by very high levels of (physiological) arousal which may have been disfunctional for a complex evaluation task. Some of the rating quality standards conveyed to participants required that reasonably fine discriminations be made between ratees (variability) and within ratees (halo). The Yekes-Dodson Law would predict that performance on
a moderately complex task such as this one might be hurt by high participant arousal levels (see Wood and Hokanson, 1965).

Recent research on social facilitation suggests that the presence of other individuals (especially in an evaluative context) leads to high arousal and in turn to lower performance on new or complex tasks. (See Porter, Lawler and Hockman, 1975, p. 413 for a recent overview of this research). The social facilitation literature is quite relevant to the present results because ratings were made in the presence of others, some of those present in the room were directly evaluating each other, and the rating task was entirely new to many of the recruits and was of at least moderate complexity. All of these factors suggest that the combined Review and Known Standards group experienced disfunctionally high levels of arousal (due to social facilitation) which may account in part for its relatively low rating quality when compared to the single component (lower arousal) groups.

To summarize, it has been suggested that potential explanations for the obtained negative R x KS interaction include: 1) operational "slippage" within the R x KS group which led to incomplete instructions to the group, 2) potential problems with the theoretical components of accountability which were suggested and manipulated, and 3)
the possible dysfunctional consequences of the high levels of arousal (motivation) within the R x KS group. Insufficient evidence exists on which to base a definitive explanation of why the dual component group performed relatively poorly. However, it is suggested that the "operational slippage" explanation in combination with the "dysfunctionally high arousal" explanation may have produced the poor performance.

The third major unexpected result within this study was the absence of consistent differences across experimental groups in responses to the "Reactions Survey" questionnaire. Several possible explanations can be suggested. One explanation could be that within the present context the Review and Known Standards manipulations were overshadowed by more salient contextual variables. If this was the case then it would be reasonable to expect no differences across the four experimental groups. A second possible explanation could be that the theoretical rationale which projected a review effect on rater resistance and satisfaction is erroneous. This seems unlikely, however, in view of the results obtained by Knouse (1977).

A third possible explanation for the uniform rater reactions could be that raters did not respond to the questionnaire honestly. It is quite feasible that raters
did not trust the CEs who told them that responses on the questionnaires were completely confidential and would be seen only by the author. If low levels of trust did prevail, respondents may have attempted to give answers high in social desirability and abstain from expressing any true negative reactions that they felt.

A fourth possible explanation for the absence of experimental effects upon rater reactions may be traceable to defects in the questionnaire itself. It is possible that items were worded or anchored in ways that were insensitive to true differences in respondent reactions. This explanation is feasible but probably somewhat far-fetched. It may be recalled that the reliability of the questionnaire was quite acceptable. The internal consistencies of item clusters (constructs) were high (appx. .85) and test retest reliabilities of repeated items within the questionnaire averaged about .90. Nonetheless, the sensitivity of items to altered respondent emotional states cannot be judged from item reliabilities.

A fifth possible explanation may rest in the manner in which questionnaire responses were analyzed. Though the procedures used to standardize individual item weights was quite defensible, the decision to retain five constructs was rather arbitrary. It is likely that constructs which were correlated as highly as these reduced the power of the MANOVA used to test for
accountability effects. However, the small number of significant univariate effects suggests that regardless of how many rater reactions constructs were retained, the number was unlikely to notably alter the results.

In summary, five potential explanations for the absence of accountability effects upon rater reactions were proposed. Of these it would seem that the most feasible explanations are these: Other contextual variables overshadowed the effects of the accountability manipulations and/or low levels of respondent trust led to socially desirable or "safe" responses which were mostly uniform across conditions.

Potential Limitations of the Present Study

As was noted in the first chapter of this report, numerous precautions were taken to help insure the internal validity and generalizability of the obtained results. Despite these precautions some possible threats may remain as viable alternative explanations for the obtained results. However infeasible, weak or far-fetched these potential factors may be, they are presented here out of concern for complete scientific recording. Though the author does not view these factors as threatening to the obtained results or conclusions, the reader is welcome to draw his/her own conclusions.
Perhaps the major threat to the internal validity of this study can be found in the operational slippages which led to less than optimal implementation of the combined Review and Knowledge of Standards group.

As was noted earlier, the experimenter administering this group, ("R"), was ill at ease during the presentation of the experimental instructions, omitted portions of the Review manipulation, and rushed through the Knowledge of Standards manipulation. It would therefore be unfair to say that the exact manipulations given to the single component groups were fully combined within the R x KS group.

Other possible threats to internal validity were largely unavoidable due to administrative limitations. First, it was not possible for all recruits to be administered the experimental conditions at exactly the same time. There were insufficient classrooms and CEs willing to serve as trainers to have everyone meet at a single time. Therefore, one potential confound of these results would be due to time differences in the administration of accountability procedures. Though recruits meeting from 1-3 P.M. did not interact with those meeting later (from 3-5 P.M.) it is possible that the time difference could have had some slight influence on recruits levels of alertness and/or motivation.
Also related to the time factor was the fact that recruits in the 3-5 P.M. groups saw the author prior to making ratings while the 1-3 P.M. groups did not. If seeing the author altered recruits perceptions of the importance of or use of the ratings being made, then the time confound may have been significant. Again, however, the author was present for fewer than 30 seconds and a major effect due to such an apparently weak and inadvertent intervention seems improbable.

Another unavoidable potential confound can be traced to the CEs who administered the experimental manipulations. It may be recalled that CE "R" administered the R x KS group at 1 P.M. and the "Review only" group at 3 P.M.; in the meantime CE "M" administered the control group at 1 P.M. and the Known Standards only group at 3 P.M. If differences in the quality or content of presentations by the two CEs were present and notable, these differences could potentially explain the obtained results. CE differences were probably not of any importance, however. Both CEs were quite verbally fluent and able to communicate equally effectively. Furthermore, experimenter differences have seldom been reported in previous studies (see Warmke & Billings, 1979).

Thus several possible threats to the internal validity of this study can be identified. These included
the content, completeness, and quality of the instructions presented, and the possible confounding due to time of day, seeing the author, and differences between CE experimenters. None of these alternative explanations appear to present strong threats to the internal validity of the present study, and certainly they are relatively minor when compared to the numerous threats within quasi-experiments. The potential threats which were suggested would have had to be strong "interventions" in their own right to serve as viable alternative explanations of the obtained results.

Potential threats to the external validity of this experiment should also be noted. These threats can involve poor generalizability due to subjects, setting, independent variable manipulations and dependent measures. The critical reviewer will note that though all of these threats are possible within any study, the strength of these threats is the critical issue. Perhaps the most important question involves what subjects, settings, and independent or dependent variables the researcher wishes to generalize the results to.

Numerous precautions were taken within this study for the express purpose of increasing generalizability. It may be recalled that 1) workers within an actual work organization served as subjects (as opposed to college students), 2) instructions and experimental conditions were administered by actual supervisors (as opposed to the
experimenter), 3) ratees were actual peers well known to the raters (as opposed to "hypothetical paper people" or videotaped ratees unknown to raters), 4) a rating scale feasible for continued use was utilized (as opposed to a "research" or "theory-oriented" form), and 5) the ratings were to be used for operational feedback and possible administrative decision-making (as opposed to use for "confidential research purposes").

Most of the above precautions are quite rare within the peer evaluation literature. The generalizability of most previous studies can be challenged on two or more of the above counts.

Even the numerous safeguards taken to insure this study's generalizability do not permit generalizing these results to any subjects, any setting, or any operationalization of the independent variables. The representativeness of the present subjects and setting must be considered. Just how "unique" or "typical" were the present police recruits and the present training setting? Were these subjects quite similar to other employees within other types of work organizations? Clearly the answers to these questions must be somewhat subjective. However, the ability to answer these questions depends to no small extent upon a clear understanding of exactly what the present subjects and setting were like.
In the interest of clarifying issues of generalizability for future researchers, a detailed description of the present setting and subjects is presented in Appendix J.

Appendix J noted that in many ways the present site was typical of other large urban police academies. The present department was more or less typical in its selection devices, in its pre-induction anticipatory socialization, and in the sorts of behaviors that were rewarded at the academy (i.e. "staying low, and keeping out of trouble").

Despite being typical in many ways the present site was atypical of police training academies in several respects. These included 1) the large discrepancies between whites and blacks in their "qualifications" and the duration of their waits prior to entering the academy, 2) the very receptive environment of the present site toward research, and 3) the relatively loose and infrequent contacts between recruits and their first level supervisors.

When the present site is compared to sites other than police training academies it may differ in numerous ways. The researcher assessing the generalizability of the present results should consider these factors and judge their relative importance on an individual basis. It is hoped that the detailed description of the present
site in Appendix J may aid future researchers to better understand the complex contextual factors which may have importantly influenced the results obtained here.

The extent to which the present ratings are generalizable to "typical" peer ratings by employees within non-training settings remains an open question. However, the precautions taken to increase the external validity of these results to other training settings should also increase the generalizability to still other types of settings and subjects. Clearly, the level of external validity of these results will depend largely on the similarity between this particular setting and subjects and those to which other researchers may wish to generalize.

Conclusions and Implications for Future Research

It has been noted that within the present study that accountability procedures did tend to affect the psychometric quality of peer ratings, yet failed to alter rater reactions. Nonetheless, a more important finding of the present study was the uniformly high quality of ratings within all rating groups. This uniform high quality may have been due to the fact that moderately high levels of Knowledge of Standards and perceptions that ratings would be reviewed (collectively if not individually) were present in all groups.
In addition to the largely contextual factors which were consistent across groups, another possible explanation for the high quality of these ratings is their source. Lawler (1967) and Kane and Lawler (1978) have argued that peers are an excellent source of ratings because of their close contacts with one another and their familiarity with performance related behavior. The peer source of the present ratings likely contributed strongly to a high capability to rate accurately, while the contextual factors noted earlier probably contributed to the high levels of motivation to rate accurately. This unusual combination of capability and motivation are seen as largely responsible for the high quality of ratings obtained.

Two major conclusions have emerged from the present study. First, accountability procedures did improve the psychometric quality (including interrater agreement and accuracy) of these ratings. Second, and probably more important was the uniformly high quality of ratings and neutral rater reactions across all experimental groups. Some of the contextual factors of the present setting may have been responsible for these results, and should not be overlooked by future researchers.

The issue of the utility of these and other possible peer ratings merits mention. The term utility has been used here to include only the psychometric quality of
ratings and user subjective reactions toward the evaluation system. However, other considerations may also be included within the utility concept. Selection test theorists like Cronbach and Gleser have used the term to include such notions as selection ratio and the base rate of successful performance in addition to test validity.

Within a performance appraisal realm the analogy would be for organizations to consider what benefits peer ratings may bring above and beyond those of the existing (supervisory) appraisal system. Consideration could be given to 1) what existing appraisals are used for (what benefits they bring), 2) the additional benefits anticipated from peer ratings, and 3) the additional costs of using peer ratings as a supplement to existing supervisory ratings.

This expanded costs-benefits utility analysis can be applied to the present setting. Doing so suggests that even though existing supervisory appraisals are used only to provide developmental feedback to recruits (benefits), the additional high quality feedback and suggestions available from well acquainted peers would probably outweigh the costs involved.

Other organizations would be well advised to consider the costs and benefits carefully before adopting a peer rating system. The costs and benefits of the system could be altered in as yet unknown ways if the peer ratings
were to be used for administrative decision-making over an extended period of time. Some factors which could enter into an organization's utility equation have not been adequately explored by existing research. These factors include the effects of the purpose of the peer ratings (feedback vs. administrative) or the long term effects of their use upon both psychometric quality and user reactions. Because the effects of some variables upon the long-term utility of peer ratings is poorly understood, practitioners should be advised to proceed cautiously.

In spite of present gaps in our understanding of the dynamics of peer ratings, the results of this project and other related research do suggest at least seven recommendations for practitioners considering the adoption of peer rating systems. First, the goal of the system should be thoroughly considered. If aid in administrative decision-making is the sole purpose, peer and/or supervisory rankings or nominations may be a more efficient and less costly method (see Kane and Lawler, 1978). If feedback for developmental purposes is desired, peer ratings may be most effective.

Second, support for the purpose of the system should be obtained from everyone who will be involved with it. Warmke & Billings (1979), for example, found that letting future raters participate in the development of the future
system is beneficial in terms of both rating quality and rater acceptance of the system. Obtaining management and supervisory support may also be critical. For example, supervisors may initially view peer ratings as a challenge to their own evaluative role.

A third consideration is the quality of the peer rating scale itself. It is suggested that behaviorally anchored scales be used and that these anchors be obtained from and scaled by those who will later use the system.

A fourth consideration is the training of raters. The present study and numerous others support the notion that clear behaviorally referenced standards of rating quality should be supplied so that raters are fully aware of the kinds of rating errors they might be prone towards. In addition to mentioning rating errors and how to avoid these, Spool (1978) has noted that effective training should provide opportunities to practice the desired behaviors and feedback to raters on how well they are doing.

A fifth consideration is the selection of raters to be used. The work of Wherry and others suggests that only raters who have had a reasonable opportunity to observe performance-related behavior should be used. This factor may suggest that peer ratings are simply not feasible within some settings.
A sixth consideration involves the procedures to be used to feed the rating results back to ratees. Will this be handled by supervisors, by peers, by some neutral outsider or will feedback be delivered at all? Supervisors would appear to be the most natural source of peer rating feedback, but if this is to be the case, perhaps some training in feedback interviewing should be provided.

Finally, the attitudes of raters toward the system should be monitored and responded to. In a setting where peer ratings are to be used over a period of time, peer raters will hear about the system before they make use of it. If what is heard is negative (negative feedback loops) then the system is less likely to serve its desired function. Avoiding negative feedback loops may only be possible if those who administer the system are responsive to negative feedback and complaints from users.

The present findings reinforce earlier suggestions by Warmke (1977) and Warmke and Billings (1979) regarding promising areas for future research. One clear message of the current study is that under proper circumstances peer ratings made for administrative as well as feedback purposes can be of very high quality. This conclusion directly contradicts the suggestion by Denisi and Mitchell (1978) that peer ratings not be used as criterion measures. Denisi and Mitchell had noted the
Several facts seem clear. Although peer ratings have been shown to be valid predictors of performance, some problems connected with their use cannot be ignored. Since most of these problems potentially would be magnified as the stakes increased, it would seem unwise to use peer ratings as criterion measures. (p. 372).

In their critique of peer ratings as criterion measures (for administrative decision making) Denisi and Mitchell emphasized rating distortions associated with friendship, subgroup effects, and stereotypes. However, Wherry (Note 5) has noted that these types of rating distortions tend to cancel one another out when several raters evaluate a given person. Therefore, friendship may increase a given ratee's evaluation from some raters but animosity may reduce the same person's evaluations from the other raters. The point is that the average rating a given ratee receives from several qualified raters tends to be a very accurate reflection of their performance (in the present study each recruit was evaluated by 9 peers). Wherry (Note 5) has presented a thorough mathematical rationale for why this will be true.

Thus, the major suggestion of the present study is that future researchers and practitioners pay more attention to peer ratings to be used not only for prediction but also for feedback purposes and/or administrative decision making.
A second suggestion is that the concept of accountability, and its components of knowledge of standards and review, be investigated further.

The present results suggest that the two components may not combine in a positive way; however, problems within the dual component condition used here prevent any solid conclusion or explanation. The success of the review only condition in increasing accuracy and interrater agreement suggest that this experimental procedure deserves further attention (both alone and in combination with Knowledge of Standards).

A third suggestion reinforced by the present findings is that the dependent variables used in performance appraisal research merit greater attention. The wide differences in results with respect to the "visible" (distributional assumption) measures and the convergent validity measures suggests that typical measures of halo, variability and leniency may not always reflect greater accuracy. Future researchers should expend more energy in attempts to measure accuracy within field settings and should use their creativity to devise new rating quality measures (such as the number of written comments and suggestions measures used here).

A fourth suggestion is that researchers and practitioners devote considerably more attention to how peers and other rating sources react to the evaluation
process. Denisi and Mitchell (1978) and Kane and Lawler (1978) have both noted the absence of research dealing with this important topic. The present study has only begun to explore the concept of rater reactions and to suggest five possible constructs underlying the process. However, a great deal of work remains to be done in this area.

A fifth suggestion is that performance evaluation studies look at the effects and/or utility of the system from a longitudinal perspective. Only longitudinal follow-ups of systems such as the peer rating system developed here will reveal whether they continue to produce the desired levels of rating quality and participant acceptance.

A sixth and final suggestion of the present research is that more attention be directed toward contextual factors within an evaluation setting which may importantly influence both rating quality and rater reactions. Warmke and Billings (1979) have already expressed the potential importance of these variables:

Contextual factors such as performance pay contingencies, political and union pressures on raters, turnover rates, time constraints, and the need to justify ratings may be more important than training in influencing the motivation of raters.

The results of the present study strongly reinforce the need to investigate contextual factors
further. If contextual factors are important, effective interventions by practitioners and researchers may need to be accompanied by structural interventions or the nature of the interventions themselves may need to be changed.

Within the present setting important contextual factors included the high level of support by management of the peer rating system, and the proposed uses of the ratings. Thus numerous contextual variables may operate differently within diverse settings. It is suggested that contextual factors be identified and measured so that these may also be experimentally manipulated in the future.

In summary, several areas are seen at meriting further attention by practitioners and researchers. These include 1) the use of peer ratings for feedback and administrative purposes, 2) the accountability concept within the performance appraisal domain, 3) the interrelations among dependent variables used to measure rating quality, 4) the reactions of raters to the evaluation process, 5) longitudinal investigation of evaluation system effectiveness, and 6) the contextual factors within organizations which may importantly influence the overall utility of ratings.
PROCESS MODEL OF PERFORMANCE RATING

RATEE CHARACTERISTICS

ORGANIZATIONAL CHARACTERISTICS

RATING INSTRUMENT

RATING PROCESS

DATA ANALYSIS

PERFORMANCE DESCRIPTION

PERSONNEL ACTION

SCALE DEVELOPMENT

POSITION CHARACTERISTICS

RATER CHARACTERISTICS
Appendix B

Acquaintance Survey
ACQUAINTANCE SURVEY

Instructions:

This questionnaire is intended to help us to better understand the interactions among recruits within a Training Academy. Your answers will be completely confidential and will not be available to Academy staff members or other employees of the department. This survey asks questions which are intended to find out which other recruits you know best. For example, you will be asked which recruits 1) you know best, 2) you could give the first name of if given their last name, and 3) you could pick out of a "lineup" if given their name. All of these questions are intended to get at how well you know particular recruits. This information may be used to improve the quality of interactions in future recruit classes. Please answer these questions as honestly as you can.
Directions: Please write the names of the five (5) recruits who you know best (not necessarily the ones you like best).

1) __________________________ 4) __________________________
2) __________________________ 5) __________________________
3) __________________________

Please write the names of the five (5) recruits you are most friendly with?

1) __________________________ 4) __________________________
2) __________________________ 5) __________________________
3) __________________________

On the next page you will be asked to place check marks next to the names of recruits who fit certain requirements. A sample is listed below.

<table>
<thead>
<tr>
<th>Recruit name</th>
<th>Given only this persons name I could pick them out of a &quot;police lineup&quot;</th>
<th>Given only this persons last name I could tell you their first name</th>
<th>Included all recruits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) John Brown</td>
<td>1)</td>
<td>1)</td>
<td>✅</td>
</tr>
<tr>
<td>2) George Crenshaw</td>
<td>2)</td>
<td>2)</td>
<td></td>
</tr>
<tr>
<td>3) Willie Miller</td>
<td>3)</td>
<td>3)</td>
<td></td>
</tr>
<tr>
<td>4) Lynn Swann</td>
<td>4)</td>
<td>✅</td>
<td></td>
</tr>
</tbody>
</table>

If you did not know the four recruits whose names were listed very well you might answer as the sample person did above. If you knew that Mr. Brown's first name was John and that Mr. Swann's first name was Lynn, but did not know Mr. Killers or Mr. Crenshaw's first names yet, you might have answered with the two checks in the right column above.

If you only knew Lynn Swann well enough to identify him by appearance alone, then you would place a check next to his name only in the left column.
Please indicate how well you know each of the recruits listed below by circling the appropriate number to the right of each recruit's name. Higher numbers indicate higher levels of knowing the recruits. Circling the number 1 would indicate not knowing the recruit at all, and circling the number 7 would indicate knowing the recruit extremely well. Your answers will be confidential and may help to stimulate improvements in present academy policies.

<table>
<thead>
<tr>
<th>Recruit Names</th>
<th>How well do you know this recruit?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recruit #1</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #2</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #3</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #4</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #5</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #6</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #7</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #8</td>
<td>not moderately extremely well</td>
</tr>
<tr>
<td></td>
<td>at all</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Recruit #</td>
<td>How well do you know this recruit?</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td></td>
<td>not at all well moderately extremely well</td>
</tr>
<tr>
<td>51</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

Please look back over your answers to see whether you have responded to all items before turning this in. This will prevent having to be re-contacted to fill in missing information. Thanks very much for your participation in this research effort.

Sincerely,

Dennis L. Harrke
The Ohio State University
APPENDIX C

Outlines of Experimental

Instructions Given to CE's
General outline of procedures: Updated version

Friday, March 9, 1979, 10m

10m-Module Room #1 (left)

Group Designation: Identified RATINGS with Training

Captain Intro comments:

Meal: General Introductory Content (11:30-12:00)

a) Peer RATINGS are being widely used (Police & Military)

b) Peer RATINGS are apparently valid because of greater exposure of peers to performance.

c) The primary PURPOSE of these RATINGS will be to provide each of you with useful FEEDBACK. In addition, these RATINGS will become a part of your ACADEMY FILE and will be available to DECISION-MAKERS and supervisors. It is hard to tell how much WEIGHT will be given to these peer RATINGS because they have not been used before.

d) ALL RATINGS will be taken using DIFFERENT PROCEDURES, with each of four groups in turn out which PROCEDURES are most effective. Mr. Warke from Ohio State will be doing some research to find out which procedures work best.

e) After you complete your evaluations, they will be TURNED IN TO YOUR CE who will summarize the RATINGS you receive in an anonymous form and give you feedback regarding the RATINGS you receive.

f) The recruits you evaluate will not see the RATINGS you gave them nor who evaluated them.

Please REMOVE & LOOK OVER THE CONTENTS of your PACKET.

Note: Instructions:
- RATES of 9 recruits to evaluate.
- RATING FORM.
- REACTIONS ALIGNED.

NOTE: Reactions survey for research purposes, will be removed from the packet by FR. Warke before the CE's see them.

h) NOTE THE CONSTRAINTS AND HIGHLIGHT IMPORTANCE OF GIVING EACH RATE A NUMERICAL RATING, KEEP TRACK OF THE TIME.
In addition to the self-development purpose of these ratings which was mentioned earlier, another purpose is to help us (the Ce's) to assess how good each of you is at observing, remembering and evaluating your fellow recruits.

For this reason each set of ratings which you turn in today will be looked over by your own Ce and evaluated to see that the feedback and ratings are accurate.

In order to be a good police officer a recruit must develop refined observational, memory and judgement skills. Each of these skills is required in order to provide useful feedback to fellow recruits.

We think it is important that these skills be developed and evaluated in some way during your stay at the academy.

After we evaluate the quality of the peer evaluations you give we will give you some feedback on your own skills as an evaluator. Your judgement and evaluation skills are important not only to your overall performance as a police officer, but also to your potential as a future supervisor where you would be required to evaluate other officers on a continuous basis.

Remember, the quality of the ratings and feedback you provide will be reviewed and evaluated objectively. So try to be as consistent and accurate as possible.

[Handwritten note: The conclusion of the feedback information is not clear due to handwriting difficulties.]
Training Condition (1 pm only)

Distribute copies of the handout "Checklist of points to remember when making evaluations".

Critical

Note that the quality of the ratings each recruit makes will be evaluated in terms of how closely the ratings coincide with the points listed on the handout.

Read over the handout, give examples, answer questions.

No Training Condition (3 pm only)

Do not distribute a handout or discuss standards of rating quality.

(1 pm-3 pm)

Final general instructions prior to rating:

✓ Please read the instructions on the rating form and try to follow these as best you can in the time you have.

"You will have about 60 minutes to finish making complete ratings of 9 of your fellow recruits.

The rating forms you fill out are 36 pages long. Try to spend no more than 2 minutes or so per page of the rating form.

✓ Be sure to complete the entire rating form before 2:30 pm.

✓ Try to make your ratings as accurately and objectively as you can.

✓ If there are any questions about the rating form or the reaction's survey, please raise your hand.
General outline of procedures: Updated version 134

Friday, March 9, 1979, 10m

New Module Room 2

General Introductory Contents (1 and 3 PM)

a) Peer ratings are being widely used (Police & Military)

b) Peer ratings are apparently valid because of greater exposure of peers to performance

c) The primary purpose of these ratings will be to provide feedback.
   In addition, the average ratings you receive will become a part of your academy FILE and will be available to DECISION-MAKERS and superintendents.
   It is hard to tell how much weight will be given to these peer ratings because they have not been used here before.

d) All recruits will be rating peer evaluations, and we are trying out slightly different procedures, with each of four groups to find out which procedures are most effective.
   Dr. Winkle from Ohio State will be doing some research to help evaluate these procedures.

e) After you complete your evaluations, they will be turned in to your CE who will summarize the ratings you receive in a non-anonymous form and give you feedback regarding the ratings you receive.

f) The recruits you evaluate will not see the ratings you gave them or know who evaluated them.

g) Please remove & lock over the contents of your packet.
   Note: Instructions may vary to evaluate RATING FORM
   READ EXAMPLES ALCUD
   NOTE: Reactions Survey for research purposes, will be removed from the packet by your CE before the CE's see them.

h) Note the constraints and emphasize importance of giving each ratee a numerical rating, keeping track of the time.

(over)

You will be making these ratings anonymously.

5. Please remove the tape with your name on it from your packet.

Even your CE will not know what ratings you made
TRAINING CONDITION: 3 PM ONLY

Distribute copies of the handout “Checklist of points to remember when giving evaluations”.

CRITICAL

Note that the quality of the ratings will be evaluated in terms of how closely the ratings coincide with the points listed on the handout.

Read over the handout, give examples, answer questions.

NO TRAINING CONDITION: 1 PM

Do not distribute a handout or discuss standards of rating quality.

FINAL GENERAL INSTRUCTIONS PRIOR TO RATING:

✓ Please read the instructions on the rating form and try to follow these as best you can in the time you have.

✓ You will have about 60 minutes to finish taking complete ratings of nine of your fellow raters.

✓ The rating forms you fill out are 36 pages long. Try to spend no more than 2 minutes or so per page of the rating form.

✓ Be sure to complete the entire rating form by 1 PM.

✓ Try to take your ratings as accurately and objectively as you can.

✓ If there are any questions about the rating form or the reaction's survey, please raise your hand.
APPENDIX D

Instructions by Captain to Recruits
Introduce Dennis

Indicate Dennis' role in working toward improving academy, tests, instructional objectives, and performance feedback systems.

Indicate Dennis' need for help & cooperation in collecting some research information for his PhD degree.

Indicate the importance of the research to the Academy and the Department.

I believe that the research involves three major aspects:

1. The objective of the research is to determine the effectiveness of the current approach to evaluating the Academy's performance.

2. The research will provide insights into the factors affecting the Academy's performance and the potential areas for improvement.

3. The research will be used to develop a comprehensive feedback system that will help the Academy identify strengths and weaknesses and take appropriate actions to improve its performance.

I would like to keep participating in the research voluntary. If you have any questions, please ask Dennis to explain things to you more thoroughly.

Dennis Johnson is question to have, and then structure the
Academy Importance.

CUTLINE OF STRATEGIC QUESTIONNAIRE (3rd to 4th for 9-5)
APPENDIX E

Checklist of Points to Remember
When Making Evaluations
CHECKLIST OF POINTS TO BE CONSIDERED IN MAKING EVALUATIONS

1) Good ratings should behaviorally describe typical present performance, and suggest ways for recruits to improve.

1a) This means, give as much of a description of each recruit's present behavior as time permits (two or more specific examples)

1b) Suggest as many ways to improve as time permits. (one or more suggestions per one of the evaluation form).

2) Good ratings should reflect the real differences among recruits in overall performance.

Each of the 5O tests at the academy includes widely varying levels of performance, and your ratings should reflect these differences. This right means:

2a) In any group of 9 recruits, there should be 6 or more numerical points' difference between the upper and lower performing recruits.

2b) Ratings should not be "bunched" all in the upper or middle portions of the scale. All or most of the points on the scale can be appropriate to use to realistically denote the recruit differences.

3) Good ratings should realistically point out individual's own personal strengths and weaknesses. It is well known that everyone has some relative strengths and weaknesses. Good ratings should take this clear.

3a) Very few recruits are really "good at everything" or "lousy at everything", so be sure that your ratings don't suggest that this is the case.

3b) On average, there should be 3 or more points' difference between the highest and lowest ratings that you give to one single recruit.

3c) Usually even a very good recruit should not be rated higher on his/her lowest dimension than another less competent recruit is rated on their highest dimension. Even top performing recruits are very good in some areas, and very good recruits are weak in some areas.

4) Try to base ratings on performance rather than on who you like or who you don't like, or the one that signifies to you. It is only human nature to allow these factors to skew your ratings to some extent. However, it is possible to minimize these biases if specific behaviors and performance are emphasized.

AS COUNSELOR EVALUATING, OR AS A QUESTION, "ARE YOU AIDING OR HARMING?", WE SUGGEST THAT EXPRESSING YOUR OPINION WILL INEVITABLY LEAD TO OR RESULT IN UNFAIR RATING.

Signed, Gay Miller, Bill Rubin, Jim Reeder, Terry Daniels,
James Smith, Ken Robinson,
Evaluators

COLUMBUS, OHIO DIVISION OF POLICE
POLICE ACADEMY

Instructions and Behavioral Examinations
for the
Personnel Evaluation and Development Form
for
Police Recruits

C) Burcin, 1977, Laydy & Ferr, 1975
Revised, 1979
The purpose of the Personnel Evaluation and Development Form for Police Recruits is to provide an objective method for evaluating the performance of recruit officers at the Police Academy and systematically communicate the performance evaluation to recruits.

The rating scales have been developed to assist you in rating the performance of recruit officers on four performance dimensions (Physical fitness, compatibility, mastery of subject matter, and leadership) which have been identified by training officers as important aspects of recruit performance at the Academy. This booklet contains rating scales and behavioral examples for each rating scale. The behavioral examples are used to assist you in making an objective judgment as to the level of performance of a recruit on the dimension being rated. The behavioral examples are the standards against which a recruit is evaluated. Therefore, in using the Form, each recruit is compared to the behavior descriptions. Recruits are not compared to one another as a part of the rating process.

The use of ratings which are based on behavioral examples has been shown to significantly reduce the subjectivity of ratings. Each rating dimension has been carefully defined and the behavioral examples have been carefully specified to insure that each rater will apply the same standards to the recruits which he/she is evaluating.

Each rating scale is composed of a vertical line of numbers from 1 to 9 with tick marks midway between the numbered lines. The behavioral examples indicate exactly which level of performance corresponds to the High, Average, and Low points on the scale. In assessing the ratings you must decide:

1) How close does the behavior of the recruit come to the examples?
2) How consistent is the behavior of the recruit?
3) What behaviors relevant to this dimension have you observed in the recruit.

Then, write down one specific example of observed behavior which supports your observations, and circle the point on the scale which most closely corresponds to the performance of the recruit. As much as possible examples used to support ratings should be specific behaviors which you personally have observed. At the bottom of each page of the Form space is provided for you to make suggestions that might help the recruit to improve his/her performance. As much as possible these suggestions should be specific things that the recruit can do to improve.

STEPS IN PREPARING AN EVALUATION FORM

Objective and valid ratings can be attained by following these steps:

1) Read the behavioral examples for the performance dimension of "Physical Fitness" on page 1 of the evaluation form.
2) To the right of these examples, in the space provided, write in a short and specific description of the recruit's recent verbal behavior on that dimension. Your examples can be similar to the behavioral examples already listed on the scale.
3) Circle the point on the numerical rating scale which best describes the level of performance you have observed and described in step 2.
4) At the bottom of each page record any constructive suggestions which might aid the recruit in improving his/her performance on the dimension.

(over)
STEPS IN FILLING AN EVALUATION FORM (cont.)

5) Read the behavioral examples for the next performance dimension.
6) Repeat steps 1 through 4 for the four performance dimensions and for each recruit you will evaluate.
7) Evaluate all recruits listed in this form as best you can. Do this even if you have doubts about how well you know them.
8) Budget your time carefully. Do not get carried away and try to spend too much time evaluating any particular recruit. You should average about two minutes per rate or about 8 minutes each for each of the 9 recruits you will be evaluating.

You will be asked to evaluate as best you can the following 9 recruits during this 80-minute evaluation session:

1. _______________________________
2. _______________________________
3. _______________________________
4. _______________________________
5. _______________________________
6. _______________________________
7. _______________________________
8. _______________________________
9. _______________________________

It is expected that you will not know some of the recruits listed above very well and that your ratings will in some cases be based upon only partial information. However, do be sure to give a numeric rating to each recruit on each dimension, regardless of your level of confidence in the rating.

Since there is somewhat limited time in which to complete these evaluations, please pace yourself continuously and keep track of the time so that you average about two minutes per rate or 8 minutes per recruit for each of the 9 recruits you evaluate.
Recruit being evaluated: ___________________________  144

Physical Fitness 1: P.E. Physical Fitness

1: P.E. PHYSICAL FITNESS: The development of physical strength, health and endurance at a superior level so that recruit can consistently handle any possible situation in police work (e.g., overtime, chase, confrontation) which requires physical fitness.

<table>
<thead>
<tr>
<th>Physical Fitness (Circle one)</th>
<th>Behavioral Indicators of High, Average, Low Performance</th>
<th>Your Description of the Recruit Relative to these Behavioral Indicators (be specific)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-</td>
<td>Participation fully during all physical training - recruits can handle level of fitness during running, swimming, calisthenics. For example, runs the mile in 7:30, can do 30 or more push-ups.</td>
<td></td>
</tr>
<tr>
<td>8-</td>
<td>Partial participation in physical training - recruits can handle level of fitness during running, swimming, calisthenics. For example, runs the mile in 7:30 or better, can do 25-30 push-ups.</td>
<td></td>
</tr>
<tr>
<td>7-</td>
<td>Limited participation in physical training - recruits can handle level of fitness during running, swimming, calisthenics. For example, runs the mile in 8:00 or better, can do 20-25 push-ups.</td>
<td></td>
</tr>
<tr>
<td>6-</td>
<td>Very limited participation in physical training - recruits can handle level of fitness during running, swimming, calisthenics. For example, runs the mile in 8:30 or better, can do 15-20 push-ups.</td>
<td></td>
</tr>
<tr>
<td>5-</td>
<td>Extremely limited participation in physical training - recruits can handle level of fitness during running, swimming, calisthenics. For example, runs the mile in 9:00 or better, can do 10-15 push-ups.</td>
<td></td>
</tr>
<tr>
<td>4-</td>
<td>Participation very limited during all physical training. This recruit may not need direct supervision. Limited participation in physical training - recruits can handle level of fitness during running, swimming, calisthenics. For example, runs the mile in 9:00 or over, and can do fewer than 10 push-ups.</td>
<td></td>
</tr>
</tbody>
</table>

SUGGESTIONS: If a grade is not given list any suggestions you may have for this recruit or the L.E.R.O.S. and/or performance in view of duties, that could be considered to be introduced in a similar or existing format.
Performance Evaluation No. 2: Compatibility

2. COMPATIBILITY: Ability to interact with fellow recruits in both social and professional relationships including accepting and giving constructive criticism, mutual decision making, and being an equal share of the work load.

<table>
<thead>
<tr>
<th>Numerical Rating</th>
<th>Excellent, above average, average, or below average</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>High description of initiate interactions with others</td>
</tr>
<tr>
<td>4</td>
<td>Good description of initiate interactions with others</td>
</tr>
<tr>
<td>3</td>
<td>Accepts by other recruits</td>
</tr>
<tr>
<td>2</td>
<td>Isolates self from others</td>
</tr>
<tr>
<td>1</td>
<td>Unwilling or hostile in all areas with others, unwilling to accept the viewpoints of others</td>
</tr>
</tbody>
</table>

To what extent could this recruit do to improve? Please be specific.
### Performance Evaluation (Primarily of Subject Matter)

**Performance Evaluation:**
Ability to learn, understand and communicate the
subject matter being taught in course.

<table>
<thead>
<tr>
<th>Numerical Rating</th>
<th>Behavioral Examples or High, Average, and Low Performance</th>
<th>Your description of the recruit presents typical performance (as specific).</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3-</td>
<td>Comments and questions, both in and out of class reflect a deep understanding of subject matter. Usually scores at or below 50% level on exams.</td>
<td></td>
</tr>
<tr>
<td>-2-</td>
<td>Comments and questions, both in and out of class reflect some interest but a good understanding of subject matter. Usually scores at or above 50% level on exams.</td>
<td></td>
</tr>
<tr>
<td>-1-</td>
<td>Comments and questions, both in and out of class reflect a lack of understanding of subject matter. Usually scores at or below 50% level on exams.</td>
<td></td>
</tr>
</tbody>
</table>

**SUGGESTIONS:**
What could this recruit do to improve? Please be specific.
Recruit being evaluated: ________________________________

Personnel Discussion "B" Leadership

**B: Leadership:** Ability to take charge of a situation and direct the behavior of others.

<table>
<thead>
<tr>
<th>Numerical Rating (circle one)</th>
<th>Behavioral examples of 1-5; Avatar: A观, and Low Performance.</th>
<th>Your description of the recruit's present typical performance (be specific).</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Assumes responsibility and effectively controls classroom in a formal leadership position.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Increases in amount above his and gets along well.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Calls for quiet when class becomes too loud.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asks questions in class.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Able to raise voice when appropriate.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SUGGESTIONS:** What could this recruit do to improve? (Circle one)

---
APPENDIX G

Reactions Survey
REACTIONS SURVEY
(for research purposes only)

[Note to Reader: mean questionnaire responses, averaged across groups, are included here so that the basic neutrality of reactions can be seen]

This Reactions Survey is a part of a research project being carried out by Dennis Varnek of the Psychology Department of the Ohio State University.

The purpose of the survey is to gather your reactions to the peer evaluation system being used by the academy, so that changes in this system can be recommended.

This survey will be removed from the manila envelope which you will turn in to your CE before your CE sees it. Your answers will therefore be completely confidential and will not be seen by anyone except Dr. Varnek. The results will be used for research purposes only in order to suggest and help bring about improvements in the system.

A summary report (which includes broad results and recommendations) will be available to interested recruits. If you would like a summary copy of the results you can indicate your interest on the form attached to the last page of this survey.

If you have any questions about this survey, please raise your hand and one of your CE’s should be able to help you.

Thanks for your help in this project.

Sincerely,

[Signature]

Dennis L. Varnek
Department of Psychology
The Ohio State University
Instructions:

In order to determine whether the instructions for the peer ratings you made were made clear to you, please respond to the following questions. For each question please circle the letter corresponding to the response you choose.

1) Which of the following were you told?
   a) Your CE will be collecting and seeing the ratings made by recruits.
   b) Your CE will not be collecting and seeing the ratings.

2) Which of the following were you told?
   a) Your ratings will not have your name on them and your CE will not know who made the ratings.
   b) Your ratings will have your name on them and your CE will know that you are responsible for them.

3) Which of the following were you told?
   a) You will be given general feedback and constructive suggestions by your CE based on the ratings and feedback given you by peers.
   b) You will be given no feedback or suggestions by anyone based on the ratings and feedback given you by peers.

4) Which of the following were you told?
   a) Since your CE will be able to tell which ratings you made, he will evaluate you on the quality of the ratings you gave.
   b) Since your CE will not be able to tell which ratings you made, he won't be able to tell how good the quality of the ratings you gave was.

5) Which of the following do you feel is most true?
   a) My CE has made very clear that he/she expects high quality ratings, and has specified some clear standards of rating quality.
   b) My CE has indicated that he/she expects high quality ratings, but has not clearly specified what is meant by quality ratings.
   c) My CE has given no indication of what he/she expects in my ratings.

6) To what extent will the quality of the ratings you gave have any impact on you or the way your CE views you?
   a) The quality of the ratings I gave will GREATLY influence the way my CE views me.
   b) The quality of the ratings I gave will MODERATELY influence the way my CE views me.
   c) The quality of the ratings I gave will influence a LITTLE the way my CE views me.
   d) The quality of the ratings I gave CANNOT influence at all the way my CE views me, because he will not know which ratings I gave.

7) Which of the following best describes your level of motivation to make accurate (fair and conscientious) ratings of your peers?
   a) Highly motivated to be accurate
   b) Somewhat motivated to be accurate
   c) Not motivated to be accurate or inaccurate
   d) Somewhat motivated to be inaccurate
   e) Highly motivated to be inaccurate
Q2 3) Which of the following best describes how CAPABLE you felt of making accurate ratings?
   a) very capable
   b) moderately capable
   c) slightly capable
   d) not capable at all

Q3 9) Which of the following best describes your present level of satisfaction with this peer rating and feedback system?
   a) very satisfied
   b) somewhat satisfied
   c) neutral (neither satisfied nor dissatisfied)
   d) somewhat dissatisfied
   e) very dissatisfied

Q4 10) Which of the following best describes your level of resistance to evaluating and providing feedback to your peers?
   a) I felt high resistance
   b) I felt moderate resistance
   c) I felt slight resistance
   d) I felt no resistance

The following questions ask that you place a check mark in the space which comes closest to describing how you felt about making peer feedback ratings. An example might be:

HAPPY _ _ UNHAPPY

Notice that there are seven blank spaces between the words happy and unhappy. If you felt closer to happy than to unhappy you might place a check-mark in one of the spaces closer to the word happy. If you felt not satisfied or felt neutral about it, you might place your check-mark exactly in the middle between the two words, and if you felt mostly unhappy you might place your check-mark closer to the word. Please place a single check mark in the space that most closely describes how you felt about the peer feedback ratings you made.

Q5 1) MOTIVATED _ _ UNMOTIVATED
Q6 2) HAPPY _ _ UNHAPPY
Q7 3) ENTHUSIASTIC _ _ UNENTHUSIASTIC
Q8 4) USEFUL _ _ USELESS
Q9 5) RELAXED _ _ TENSE
Q10 6) GOOD _ _ BAD
Q11 7) SATISFIED _ _ DISSATISFIED
Q12 8) INTERESTED _ _ BORED
Q13 9) FAIR _ _ UNFAIR
Q14 10) CONSCIENTIOUS _ _ CARELESS
Q15 11) RESISTANT _ _ NON-RESISTANT
Q16 12) CAPABLE _ _ UNCAPABLE

(continued on next page)
Please place a checkmark in the space that most closely describes how you felt about the peer feedback ratings you made.

Q17 13) ACCURATE ______ [ ] INACCURATE
Q18 14) APPREHENSIVE ______ [ ] CALM
Q19 15) WORTHWHILE ______ [ ] WORTHLESS
Q20 16) IMPORTANT ______ [ ] UNIMPORTANT
Q21 17) BIASED ______ [ ] UNBIASED
Q22 18) PLEASED ______ [ ] UPSET
Q23 19) HIGH QUALITY ______ [ ] LOW QUALITY
Q24 20) COMPETENT ______ [ ] INCOMPETENT
Q25 21) ACCEPTING ______ [ ] CRITICAL
Q26 22) DIFFICULT ______ [ ] EASY
Q27 23) FAIR ______ [ ] UNFAIR
Q28 24) TENSE ______ [ ] RELAXED
Q29 25) MOTIVATED ______ [ ] UNMOTIVATED
Q30 26) INACCURATE ______ [ ] ACCURATE

For the questions below please circle the response which comes closest to reflecting how you feel.

Q31 1) Do you feel that the feedback and ratings you made may help other recruits understand better their own individual strengths and weaknesses?
   a) Yes
   b) Undecided
   c) No

Q32 2) Do you feel that peer evaluations should be looked at when future promotional decisions are being made?
   a) Yes
   b) Undecided
   c) No

Q33 3) How do you feel about the peer rating system you have been involved in?
   a) I think the system is basically a good one
   b) I thought keeping peer evaluations but modifying the system
   c) I would prefer having no peer evaluative system at all.
   d) Undecided
Please give your reactions to the peer ratings you have made by circling the number to the right of each statement below which most nearly reflects your opinion. Responses will range from SA (Strongly Agree) through U (Undecided) to SD (Strongly Disagree).

Q35 1) I feel confident that the evaluations I made were accurate.
   SA 4 3 2 1  SD

Q36 2) I felt resistant to making the evaluations.
   SA 4 3 2 1  SD

Q37 3) I feel that recruits are in a good position to observe and evaluate other recruits on the rating dimensions that were used.
   SA 4 3 2 1  SD

Q38 4) I felt good about making the evaluations which I made.
   SA 4 3 2 1  SD

Q39 5) I feel that the Training Academy would do well to adopt a peer evaluation system like this one.
   SA 4 3 2 1  SD

Q40 6) I feel it was difficult to evaluate my peers on the evaluation forms.
   SA 4 3 2 1  SD

Q41 7) I feel that peer evaluations like the one we made can provide useful feedback to recruits.
   SA 4 3 2 1  SD

Q42 8) I felt personal involvement in evaluating recruits on the evaluation forms that were used.
   SA 4 3 2 1  SD

Would you like a copy of the Summary Report which describes this research project, its results, and its recommendations?

Yes  
No

If you would like a copy, please write your address in the space provided below. Your response will still remain confidential and will be seen only by the researcher.

Address:

__________________________________________

Thanks very much for your help in this project.
APPENDIX H-1

MEANS AND SD<sup>2</sup> OF REACTIONS SURVEY ITEM RESPONSES BY EXPERIMENTAL GROUPS, AND UNIVARIATE F-VALUES OF EFFECTS DUE TO REVIEW (R), KNOWLEDGE OF STANDARDS (KS) AND RxKS INTERACTION

<table>
<thead>
<tr>
<th>Questionnaire Section &amp; Item Numbers:</th>
<th>MEANS AND SD&lt;sup&gt;2&lt;/sup&gt; BY GROUP</th>
<th>F-VALUES OF EFFECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1 (Review &amp; Known Standards)</td>
<td>Group 2 (Review)</td>
</tr>
<tr>
<td>Q1</td>
<td>Mean</td>
<td>S.D.</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>M</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>.53</td>
</tr>
<tr>
<td>Q3</td>
<td>M</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>.89</td>
</tr>
<tr>
<td>Q4</td>
<td>M</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.48</td>
</tr>
<tr>
<td>Q5</td>
<td>M</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.16</td>
</tr>
<tr>
<td>Q6</td>
<td>M</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.41</td>
</tr>
<tr>
<td>Q7</td>
<td>M</td>
<td>4.14</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.51</td>
</tr>
<tr>
<td>Q8</td>
<td>M</td>
<td>3.29</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.54</td>
</tr>
<tr>
<td>Q10</td>
<td>M</td>
<td>3.31</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.44</td>
</tr>
<tr>
<td>Q11</td>
<td>M</td>
<td>3.71</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.59</td>
</tr>
<tr>
<td>Q12</td>
<td>M</td>
<td>2.79</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>1.12</td>
</tr>
</tbody>
</table>

*p < .10

**p < .05

***p < .01
### APPENDIX H-2

**MEANS AND SDs OF REACTIONS SURVEY ITEM RESPONSES BY EXPERIMENTAL GROUPS, AND UNIVARIATE F-VALUES OF EFFECTS DUE TO REVIEW (R), KNOWLEDGE OF STANDARDS (KS) AND RxKS INTERACTION**

<table>
<thead>
<tr>
<th>Questionnaire Section &amp; Item Numbers</th>
<th>Group 1 (Review &amp; Known Standards)</th>
<th>Group 2 (Review)</th>
<th>Group 3 (Known Standards)</th>
<th>Group 4 (Control)</th>
<th>F-Values of Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 13</td>
<td>Mean 2.79 S.D. 1.25</td>
<td>3.85</td>
<td>3.15</td>
<td>3.23</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.77</td>
</tr>
<tr>
<td>Q 14</td>
<td>Mean 2.14 S.D. 1.03</td>
<td>1.77</td>
<td>2.15</td>
<td>2.00</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.39</td>
</tr>
<tr>
<td>Q 15</td>
<td>Mean 4.57 S.D. 1.95</td>
<td>3.70</td>
<td>4.08</td>
<td>4.62</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.18**</td>
</tr>
<tr>
<td>Q 16</td>
<td>Mean 3.57 S.D. 1.60</td>
<td>3.85</td>
<td>4.31</td>
<td>3.31</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.41</td>
</tr>
<tr>
<td>Q 17</td>
<td>Mean 2.79 S.D. 1.25</td>
<td>1.95</td>
<td>1.38</td>
<td>1.85</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.67</td>
</tr>
<tr>
<td>Q 18</td>
<td>Mean 4.50 S.D. 1.40</td>
<td>4.31</td>
<td>4.85</td>
<td>4.85</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.08</td>
</tr>
<tr>
<td>Q 19</td>
<td>Mean 3.64 S.D. 1.69</td>
<td>4.00</td>
<td>3.31</td>
<td>3.92</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.15</td>
</tr>
<tr>
<td>Q 20</td>
<td>Mean 3.43 S.D. 1.45</td>
<td>4.54</td>
<td>3.31</td>
<td>3.77</td>
<td>3.88**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.88*</td>
</tr>
<tr>
<td>Q 21</td>
<td>Mean 4.29 S.D. 1.33</td>
<td>4.54</td>
<td>4.23</td>
<td>4.58</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.01</td>
</tr>
<tr>
<td>Q 22</td>
<td>Mean 3.79 S.D. 1.53</td>
<td>3.00</td>
<td>3.46</td>
<td>3.52</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.50</td>
</tr>
<tr>
<td>Q 23</td>
<td>Mean 2.93 S.D. 1.07</td>
<td>2.77</td>
<td>3.77</td>
<td>3.25</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.13*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.26</td>
</tr>
<tr>
<td>Q 24</td>
<td>Mean 2.93 S.D. 1.00</td>
<td>3.15</td>
<td>3.46</td>
<td>3.00</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.76</td>
</tr>
</tbody>
</table>

\[c = p < .10\]
\[** = p < .01\]
**APPENDIX H-3**

**MEANS AND SDs OF REACTIONS SURVEY ITEM RESPONSES BY EXPERIMENTAL GROUPS, AND UNIVARIATE F-VALUES OF EFFECTS DUE TO REVIEW (R), KNOWLEDGE OF STANDARDS (KS) AND RxKS INTERACTION**

<table>
<thead>
<tr>
<th>Questionnaire Section &amp; Item Numbers</th>
<th>Group 1 (Review &amp; Standards)</th>
<th>Group 2 (Known Standards)</th>
<th>Group 3 (Review)</th>
<th>Group 4 (Control)</th>
<th>F-Values of Effects</th>
<th>Known Effects</th>
<th>R* Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q25</td>
<td>Mean 3.46</td>
<td>3.38</td>
<td>3.77</td>
<td>3.08</td>
<td>.00</td>
<td>1.29</td>
<td>2.44</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.51</td>
<td>1.39</td>
<td>1.30</td>
<td>1.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q26</td>
<td>Mean 4.00</td>
<td>3.77</td>
<td>3.54</td>
<td>4.42</td>
<td>.09</td>
<td>.00</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.75</td>
<td>2.05</td>
<td>1.71</td>
<td>2.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q27</td>
<td>Mean 2.86</td>
<td>2.85</td>
<td>3.08</td>
<td>3.23</td>
<td>.14</td>
<td>.32</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.29</td>
<td>2.34</td>
<td>1.66</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q28</td>
<td>Mean 4.14</td>
<td>3.62</td>
<td>4.15</td>
<td>3.62</td>
<td>.13</td>
<td>2.49</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.35</td>
<td>1.50</td>
<td>1.41</td>
<td>1.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q29</td>
<td>Mean 4.29</td>
<td>4.62</td>
<td>4.83</td>
<td>5.54</td>
<td>3.35*</td>
<td>1.07</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>S.D. 2.16</td>
<td>1.80</td>
<td>1.75</td>
<td>1.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q30</td>
<td>Mean 2.36</td>
<td>3.85</td>
<td>3.23</td>
<td>3.23</td>
<td>.46</td>
<td>.46</td>
<td>5.82*</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.01</td>
<td>1.68</td>
<td>1.36</td>
<td>2.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q31</td>
<td>Mean 5.29</td>
<td>4.70</td>
<td>4.62</td>
<td>5.23</td>
<td>.01</td>
<td>.25</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.33</td>
<td>2.29</td>
<td>1.89</td>
<td>1.64</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$c = p < .10$  
$* = p < .05$  
$** = p < .01$
### APPENDIX H-4

**Means and SDs of Reactions Survey Item Responses by Experimental Groups, and Univariate F-Values of Effects Due to Review (R), Knowledge of Standards (KS) and RxKS Interaction**

<table>
<thead>
<tr>
<th>Questionnaire Section &amp; Item Numbers*</th>
<th>Group 1 (Review &amp; Known Standards)</th>
<th>Group 2 (Review)</th>
<th>Group 3 (Known Standards)</th>
<th>Group 4 (Control)</th>
<th>F-Values of Effects</th>
<th><strong>Known Effects</strong></th>
<th><strong>R x KS Effects</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 32</td>
<td>Mean 1.57</td>
<td>2.08</td>
<td>1.62</td>
<td>1.58</td>
<td>1.12</td>
<td>.33</td>
<td>3.51*</td>
</tr>
<tr>
<td></td>
<td>S.D. .76</td>
<td>.76</td>
<td>.77</td>
<td>.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 33</td>
<td>Mean 2.21</td>
<td>2.46</td>
<td>2.15</td>
<td>1.67</td>
<td>3.47*</td>
<td>.56</td>
<td>5.00*</td>
</tr>
<tr>
<td></td>
<td>S.D. .88</td>
<td>.88</td>
<td>.80</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 34</td>
<td>Mean 2.85</td>
<td>2.61</td>
<td>2.62</td>
<td>1.67</td>
<td>5.81*</td>
<td>5.81*</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.04</td>
<td>1.18</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 35</td>
<td>Mean 4.00</td>
<td>.33</td>
<td>.69</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 36</td>
<td>Mean 2.86</td>
<td>.34</td>
<td>2.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 37</td>
<td>Mean 3.29</td>
<td>2.62</td>
<td>3.08</td>
<td>3.69</td>
<td>1.93</td>
<td>.00</td>
<td>4.18*</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.27</td>
<td>1.26</td>
<td>1.12</td>
<td>1.03</td>
<td>.07</td>
<td>.61</td>
<td>1.45</td>
</tr>
<tr>
<td>Q 38</td>
<td>Mean 3.36</td>
<td>.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 39</td>
<td>Mean 2.86</td>
<td>2.10</td>
<td>.23</td>
<td>2.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 40</td>
<td>Mean 3.21</td>
<td>.20</td>
<td>1.48</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 41</td>
<td>Mean 3.36</td>
<td>2.77</td>
<td>3.31</td>
<td>3.54</td>
<td>2.96*</td>
<td>.31</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>S.D. 1.00</td>
<td>1.48</td>
<td>1.18</td>
<td>1.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 42</td>
<td>Mean 3.42</td>
<td>.00</td>
<td>.24</td>
<td>.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX I
C INTERKATED AGREEMENT PROGRAM
DIMENSION NM=54, NPER=54, NRATE, NPER
10 CONTINUE
C READ NUMBER IN CE AND EXPERIMENTAL GROUP COMBINATION
READ*000, NPER
500 FORMAT(12)
IF(NPER.EQ.01 GO TO 900
C ZERO ARRAYS
DO 3 J=1,54
NMJ=0
DO 3 J=1,5
DO 3 K=1,5
3 XIKJ=0
C READ RATING DATA FOR LE-EXP GROUP COMBINATION
READNPER
DO 7 NMT=1
IF(NMT.EQ.501 NAME=I+1+I+1+I+1
7 FORMAT(A+6X+I2+1X+4F5.1+2X+I2+1X+4F5.1)
IF(I+EQ.O) GO TO 5
IF(1+EQ.9) GO TO 15
IF(NMAT=NMAT(K)+1
DO 11 K=1,4
11 XIKMAT(KJ)=01+N(K)
15 IF(1+EQ.9) GO TO 5
IF(KMAT(KJ)+1=01+N(K)
DO 16 K=1,4
16 XIKMAT(KJ)=01+N(K)
5 CONTINUE
NWRITE=0
IF(NWRITE(1).EQ.NPER) GO TO 500
NWRITE=NWRITE+1
DO 99 J=1,54
99 IF(J.EQ.K=1,4)
500 CONTINUE
WRITE*0(501) NWRITE(NPER+1)
501 FORMAT(16X+I2+5X+C15)
IF(NWRITE.LT.J+1) GO TO 900
DO 505 J=1,NPER
505 WRITE=16+5C2)
(1Y(1+K+J+1)K=1+1+K=1+NPER)
506 FORMAT(15X+C15)
C COMPUTE FAIR AVERAGE CORRELATIONS
DO 600 IF=1,4
IF(NPER.LT.1) GO TO 600
NPER
NEO=0
600 CONTINUE
TCCC=CC
31 IF(N=NPER) GO TO 4C
NE=1
DO 32 =1,4
3 CALL CORR(1+K+1+Y(+4+1+1+CORR,NRATE)
IF(TCORNE=5+30 GO TO 32
NCOR=NCOR+1
TCOR=TCOR+CC
32 CONTINUE
NE=1
GO TO 31
}
600 CONTINUE
   FORMAT(11X,1X,A20,1X,5F9.4,1X,4F10.6,5X,15)
   CONTINUE
   FORMAT(11X,1X,A20,1X,5F9.4,1X,4F10.6,5X,15)
   GO TO 10
   CONTINUE
   DO 10 K=1,N
      YSU=YSU+Y(K)
      XSU=XSU+X(K)
   CONTINUE
   CONTINUE
   IF(TAY<YSU) RETURN
   CLR4=(FINAY-YSU)/SQR(T4)
   RETURN
   END
Accuracy Program (contd.)

IF(K.EQ.K2) GO TO 40
NPER2*PER2+1
CALL CONTINUE

IF(K.EQ.K2) GO TO 40
NPER2*PER2+1
CALL CONTINUE

502 FORMAT(4X,4X,3F10.3)
GO TO 20
CONTINUE
STOP
END
APPENDIX J
APPENDIX J

Description of Columbus, Ohio Police Department and Academy

The present police academy is the primary source of new police officers for the Columbus, Ohio, metro area. The 1979 edition of the Rand McNally Road Atlas lists the population of Columbus as 582,000. The metro area more nearly approximates 900,000. The Columbus police department employs over 1,000 personnel. The reader interested in a microscopic view of some of the features of this particular department is referred to a large scale survey research project which examined many of the department's operating features (see Jermier and Berkes, 1979). The present department is in many ways typical of any large urban police department and thus mirrors most of the "contextual" factors described by Van Maanen (1975, p. 209).

The officer selection process is similar to that of many other large departments. The applicant must pass a Civil Service written exam, background investigation (including polygraph), medical exam, and oral interview. Recruits who pass all of these exams are asked to wait until a police academy class can accommodate them.
The police department in Columbus has been under a court imposed minority hiring quota since 1977. The department was thus required to hire 30% black officers into the present academy class. This quota, and the low availability of qualified black applicants led to a situation where the typical white recruit was considerably more qualified than the typical black recruit. In fact, a clear double standard was used in hiring depending upon race. For example, the lowest Civil Service test score by a white recruit was above the highest score of any of the present classe's black recruits.

Though the academy is not viewed as an adjacent to the selection process the poor qualifications of the black recruits led to six blacks (and no whites) failing out of the academy due to low in-class test scores. The researcher personally reviewed the test items used and found them to be both content valid and job related. The tests did not appear in any way to be racially biased.

In addition to the racial differences in academic qualifications the court imposed quota also led to a large discrepancy between blacks and whites in the amount of time an applicant had to wait before being hired. Some white recruits had waited as long as three to four years before beginning at the academy; whereas black recruits typically had to wait only until the next scheduled recruit class started (usually every 6 months). Thus there
existed differences between blacks and whites in academic qualifications and in the average wait prior to academy induction.

Van Maanen (1975) has noted that in police organizations the protracted and often arduous screening and waiting process leads to a lengthy state of "anticipatory socialization" by applicants and "assures that those who join the organization will have strong positive attitudes concerning their new jobs" (p. 221). Van Maanen (1975) goes on to describe nicely the police academy experience. He notes the often arbitrary discipline and the social norms which often favor unquestioning obedience to authority, staying low, and avoiding trouble (p. 222). The police academy experience involves considerable adjustment and changes in attitudes. Jermier & Berkes (1979) have specifically looked at the Columbus Police Department and studied its typical form of "Police Command Bureaucracy."

Berkes (Note 9), who was engaged in a participant-observation study of the Columbus Police Academy at the time of this study, has noted that the present setting was characterized by relatively loose supervision. Contact between recruits and their immediate supervisors (CEs) was sporadic and infrequent. The 6 CEs taught different courses at the academy and were in contact with recruits primarily when they were teaching. Outside of in-class
teaching, GEs met with recruits primarily to discipline or tutor them if problems arose.

The present police academy may have been atypical in the sense that it seems to have been highly receptive toward research. At least two other research projects were being carried out by separate researchers in addition to the present project. One of these (the participant observation study by Berkes) was quite obtrusive, and the present study too required considerable adjustment on the part of the academy schedule. Thus the present academy may be different from those which turned down Van Maanen's requests for a research site. He noted that 20 departments were contacted before one agreed to his unobtrusive questionnaire project.

The typical recruit at the academy is difficult to describe. However, the typical white recruit was male, about 25 years old and had some college education. The typical black recruit was about 22 years old, either male or female, and had only a high school education.
**Note:** Women are graded higher than men; i.e., 1.5 mile run.

<table>
<thead>
<tr>
<th>Fitness Category</th>
<th>Agility Run (sec)</th>
<th>1,5 Mile Run (min)</th>
<th>Body Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOMAN</td>
<td>Below 171</td>
<td>Below 16:1</td>
<td>Below 18:19</td>
</tr>
<tr>
<td></td>
<td>Good</td>
<td>16:2</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>17:8</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>Below Average</td>
<td>18:7</td>
<td>20:2</td>
</tr>
<tr>
<td></td>
<td>Above</td>
<td>20:3</td>
<td></td>
</tr>
</tbody>
</table>

**Average**

12:01 to 14:31

<table>
<thead>
<tr>
<th>Fitness Category</th>
<th>Flexibility (in.)</th>
<th>Agility Run (sec)</th>
<th>1,5 Mile Run (min)</th>
<th>Body Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>Above 25.9</td>
<td>Above 43</td>
<td>Above 50</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>26.8</td>
<td>42</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>19.6</td>
<td>28</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Below Average</td>
<td>16.5</td>
<td>19</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>10.4</td>
<td>Below 4</td>
<td>Below 23</td>
<td></td>
</tr>
</tbody>
</table>

**Fitness Category**

<table>
<thead>
<tr>
<th>Category</th>
<th>Resting Heart Rate</th>
<th>Treadmill Time</th>
<th>Maximum Oxygen Intake</th>
<th>Step Test</th>
<th>12-Min Run</th>
<th>Body Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>14</td>
<td>Above 13:15</td>
<td>Above 52.9</td>
<td>Below 10:15</td>
<td>Above 10:15</td>
<td>Below 18:19</td>
</tr>
<tr>
<td>Good</td>
<td>15</td>
<td>13:11</td>
<td>52.8</td>
<td>10:16</td>
<td>1:74</td>
<td>6.3</td>
</tr>
<tr>
<td>Average</td>
<td>11:24</td>
<td>44.8</td>
<td>12:01</td>
<td>1:50</td>
<td>17.7</td>
<td></td>
</tr>
<tr>
<td>Below Average</td>
<td>11:23</td>
<td>44.7</td>
<td>12:01</td>
<td>1:49</td>
<td>17.7</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>10:30</td>
<td>40.8</td>
<td>14:30</td>
<td>1:25</td>
<td>22.6</td>
<td></td>
</tr>
</tbody>
</table>

**Very Poor**

<table>
<thead>
<tr>
<th>Category</th>
<th>Resting Heart Rate</th>
<th>Treadmill Time</th>
<th>Maximum Oxygen Intake</th>
<th>Step Test</th>
<th>12-Min Run</th>
<th>Body Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>8:41</td>
<td>32.8</td>
<td>16:31</td>
<td>1:24</td>
<td>22.7</td>
<td>33.3</td>
</tr>
</tbody>
</table>
The measurement of total body strength is very difficult since there are so many different muscles and muscle groups which would have to be included. A number of individual strength tests correlate well with the average strength as measured by numerous muscle groups. One of these is hand grip strength, and it can be said that it fairly representative of total body strength.

Muscular Strength
- is measured by a hand grip dynamometer.
- is valued as an indicator of physical fitness.
- decreases with inactivity.
- increases with weight.
- increases with height.
- decreases rapidly to age 60, then increases to age 30.
- is influenced by emotions and illnesses.

### Classification

<table>
<thead>
<tr>
<th>CLASSIFICATION</th>
<th>RIGHT GRIP (lb.)</th>
<th>LEFT GRIP (lb.)</th>
<th>TOTAL (lb.)</th>
<th>TOTAL (kg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXCELLENT</td>
<td>164</td>
<td>162</td>
<td>326</td>
<td>148.2</td>
</tr>
<tr>
<td>VERY GOOD</td>
<td>159</td>
<td>163</td>
<td>326</td>
<td>145.6</td>
</tr>
<tr>
<td>ABOVE AVERAGE</td>
<td>155</td>
<td>162</td>
<td>317</td>
<td>138.6</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>151</td>
<td>147</td>
<td>298</td>
<td>135.5</td>
</tr>
<tr>
<td>BELOW AVERAGE</td>
<td>146</td>
<td>142</td>
<td>288</td>
<td>129.9</td>
</tr>
<tr>
<td>POOR</td>
<td>142</td>
<td>137</td>
<td>281</td>
<td>128.0</td>
</tr>
<tr>
<td>VERY POOR</td>
<td>137</td>
<td>132</td>
<td>269</td>
<td>122.3</td>
</tr>
<tr>
<td></td>
<td>133</td>
<td>127</td>
<td>260</td>
<td>119.0</td>
</tr>
<tr>
<td></td>
<td>129</td>
<td>122</td>
<td>251</td>
<td>114.1</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>117</td>
<td>241</td>
<td>109.6</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>113</td>
<td>233</td>
<td>105.5</td>
</tr>
<tr>
<td></td>
<td>115</td>
<td>107</td>
<td>222</td>
<td>100.9</td>
</tr>
<tr>
<td></td>
<td>111</td>
<td>102</td>
<td>213</td>
<td>96.8</td>
</tr>
<tr>
<td></td>
<td>107</td>
<td>97</td>
<td>204</td>
<td>92.7</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>92</td>
<td>194</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>87</td>
<td>187</td>
<td>84.1</td>
</tr>
<tr>
<td></td>
<td>89</td>
<td>77</td>
<td>166</td>
<td>75.5</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>72</td>
<td>167</td>
<td>74.4</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>67</td>
<td>147</td>
<td>66.8</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>62</td>
<td>131</td>
<td>59.8</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>57</td>
<td>125</td>
<td>56.7</td>
</tr>
</tbody>
</table>
REFERENCE NOTES


REFERENCES


Borman, W. C. The rating of individuals in organizations: An alternative approach. Organizational Behavior and Human Performance, 1974, 12, 105-127.


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


Spool, M. D. Training Programs for Observers of Behavior: A Review.


