Spring 1987

The Influence of Rater Training, Scale Format, and Rating Justification on the Quality of Performance Ratings by Three Rater Sources

Steven B. Woods
Old Dominion University

The Influence of Rater Training, Scale Format, and Rating Justification on the Quality of Performance Ratings by Three Rater Sources

by

Steven B. Woods
B.A. June 1981, Southeastern Massachusetts University
M.S. August 1984, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

PSYCHOLOGY

OLD DOMINION UNIVERSITY
May, 1987

Approved by:

Terry L. Dickinson (Director)

DEDICATION

To Jill Woods, my best friend, and my parents, George and Ruth Woods. My parents' beliefs, guidance, and love taught me to be strong, creative, and independent. Jill has shown me that love goes beyond one's self and has continued to teach where my parents left off. I could not be in better hands.

ACKNOWLEDGEMENTS

It is with great pleasure that I take this opportunity to acknowledge the support I have received from my dissertation committee members, friends, and family to this work and to my professional development over the past five years.

I would like to thank my dissertation committee members, Drs.: Terry Dickinson, Glynn Coates, Robert McIntyre, and Michael Secunda for their support on a difficult task. In addition, thanks to Dr. Coates for your analytic support and patience in answering innumerable statistical inquires. To Dr. McIntyre, thanks for your unyielding energy. Our periodic conversations in the hallways were invaluable in lifting my spirits in times of need. Your critical eye has also served to enhance the quality of this work. Also, thanks to Dr. Secunda who taught me the politics of the real world and how to be successful in it. I am still learning but you have started me on my way.

Special thanks to Dr. Terry Dickinson who chaired my dissertation committee. Over the past few years you have been a source of insight, support, and review. I realize the added strain of completing this task over a great distance and I thank you for your patience and efforts toward this end. You have served not only as an insightful teacher, but as a mentor and friend as well. Although this work signifies the end of one phase of our relationship, I look forward to a productive and lasting friendship beyond graduate school.

In addition, I would like to single out Dr. Don Davis whose influence on my professional development dates back to my first semester of graduate training. Although we have not worked closely over the past two years, your professionalism and pursuit of excellence have left an indelible mark on me. You are my professional cornerstone and the work ethic I take with me is the product of your efforts.

I would also like to thank my friends and colleagues, in particular, John Mathieu, Scott Tannenbaum, Mark Teachout, and Todd Silverhart. John Mathieu was instrumental in my early years and taught me the "nuts and bolts" of scientific research. He passed me on to Scott Tannenbaum who is responsible for kindling my interest in the practical aspects of Human Resources and the desire to find our niche, as I/O psychologists, in the business world. I thank you both. Also, to Mark Teachout and Todd Silverhart who have experienced the stressors of graduate school with me over the past five years, I thank you for your friendship and support. It really hasn't been that bad, has it?

I would also like to acknowledge Eric Vanetti for his efforts in the data collection phase of this work. Your generous contribution of time and effort was greatly appreciated.

Lastly, and perhaps most importantly, I would like to thank Jill Woods and my family for putting up with me. You have provided friendship, love, and support during an intense period of my life. Now it's time to reap the rewards.

Table of Contents

List of Tables

List of Tables (Continued)

List of Figures

ABSTRACT

THE INFLUENCE OF RATER TRAINING, SCALE FORMAT,

AND RATING JUSTIFICATION ON THE QUALITY OF

PERFORMANCE RATINGS BY THREE RATER SOURCES

Steven B. Woods
Old Dominion University, 1987
Director: Dr. Terry L. Dickinson

Theoretical support for the use of different rater sources (e.g., self, peer, supervisor, observer) in the performance appraisal process is considerable. However, despite this evidence and the intuitive appeal of using multiple rater sources, the empirical evidence directly comparing different rater sources is both scarce and inconsistent. The primary focus of the present study was to examine systematically the influence of rater training, scale format, and rating justification on the quality (i.e., convergent and discriminant validity, halo, leniency) of ratings exhibited by three rater sources (i.e., self, peer, observer). Ninety-one undergraduate students participated in a videotaped role play exercise and returned at a later time to take part in a three-hour rating session. These individuals provided self- and peer ratings. Forty-five advanced undergraduate students participated in a similar rating session and provided observer ratings. Convergent validity, discriminant validity, and halo were tested with the multitrait-multimethod analysis of variance (MTMM ANOVA) approach. To assess the influence of training, scale format, and rating justification on the quality of

performance ratings, each experimental condition was treated as a MTMM design and separate ANOVAs were calculated. A 2 (Training) x 2 (Format) x 2 (Justification) x 3 (Rater Sources) x 4 (Dimensions) ANOVA was computed to test the effects of the experimental conditions on the leniency of performance ratings across rater sources.

Mixed support was found for the ability of these variables to influence the quality of performance ratings given by the three rater sources. Specifically, training and the use of the behavioral checklist increased discriminant validity and reduced halo, while raters who had to justify their performance ratings exhibited lower discriminant validity than raters who did not have to justify their ratings. With respect to leniency, the level of ratings across the three rater sources was affected by the variables of interest. Training and the use of the behavioral checklist helped to reduce leniency in self-ratings in those situations when raters had to justify their performance ratings.

These results lend support for the use of training and the behavioral checklist to improve the overall quality of performance ratings given by different rater sources. However, future research should assess what specific training program content is needed to improve convergent validity when the behavioral checklist is used. In addition, research must be conducted to identify which rater sources provide high-quality ratings on which performance dimensions if a multiple-method approach to the assessment of job performance is desired.

THE INFLUENCE OF RATER TRAINING, SCALE FORMAT, AND RATING JUSTIFICATION ON THE QUALITY OF PERFORMANCE RATINGS BY THREE RATER SOURCES

I. INTRODUCTION

Performance evaluation is an important component in the information and control system of most organizations. However, no one approach has proven completely satisfactory, particularly for professional employees. Performance appraisal systems are often viewed with the same enthusiasm as "income tax forms, typically described by both subordinates and supervisors as better than nothing at all" (McCall & DeVries, 1976, p. 2). Attitude surveys (e.g., DeVries & McCall, 1976) as well as informed opinion (e.g., Porter, Lawler, & Hackman, 1975) confirm this general ambivalence toward appraisal. Despite these shortcomings, surveys of managers from both large and small organizations indicate that they regard performance appraisals as an important assessment tool and are unwilling to abandon them (Zawacki & Taylor, 1976).

Formal performance appraisal systems are designed to meet three basic needs, one for the organization and two for the individual: "(1) they provide systematic judgments to back up decisions about placement, promotions, terminations, and salary increases; (2) they are a means of telling an employee how they are doing, and suggest needed changes in behaviors, attitudes, skills or job knowledge; and (3) they are also used as a basis for the coaching and counseling of

the individual by the supervisor" (McGregor, 1957, p. 89). Unfortunately, numerous authors have expressed disappointment in the lack of success organizations have experienced with most performance appraisal systems (Carroll & Schneier, 1982; Landy & Farr, 1980; McCall & DeVries, 1976). It is widely accepted that performance appraisals are prone to bias, that they do not demonstrate high levels of accuracy, and are not readily accepted by users (Banks & Roberson, 1985). Recently, attempts to overcome these difficulties have placed primary emphasis on technical issues, e.g., the advantages and disadvantages of various rating formats, sources of rating error, and problems of unreliability in performance observation and measurement (Landy & Farr, 1980; McIntyre, Smith, & Hassett, 1984). Despite some gains, these strategies have had relatively little impact on the accuracy and/or acceptance of ratings (Banks & Roberson, 1985).

One of the significant research trends in this area has dealt with the type of rater conducting the performance rating (Landy & Farr, 1980). It has been estimated that over 95% of the performance appraisals conducted at lower and middle management levels are performed by the individual's immediate supervisor (Lacho, Stearns, & Villere, 1979; Lazer & Wikstrom, 1977). There has been considerable dissatisfaction with this practice, however, because it is well documented that supervisory ratings are susceptible to intentional and unintentional bias in the rating process (e.g., Landy & Farr, 1980). As a consequence, self-, peer, subordinate, and outside observer ratings have been suggested as likely alternatives to the traditional supervisory ratings. Of these methods, peer ratings (e.g., Kraut, 1975) and self-ratings (e.g., Mabe & West, 1982) appear to have

commanded the most attention though recent work can be found on subordinate ratings (e.g., Mount, 1984; Tsui, 1983; Tsui & Ohlott, 1986).

Several advantages exist for obtaining ratings from different sources. Supervisory ratings have traditionally been included because it is assumed that the supervisor has the best overview of the situation and knows best how the incumbent's job behavior contributes to the overall goals of the organization (Lawler, 1967). Self-ratings may be used in one of several ways. They may be substituted for supervisory ratings in those situations where the supervisor does not adequately know the work performance of the incumbent. Or, they may be obtained to increase the incumbent's acceptance of any future administrative action based on the ratings. Peer evaluations, on the other hand, are relevant because peers are best situated to evaluate how the co-worker performs in terms of lateral relationships in working toward an organization's goals (Lawler, 1967). Further, empirical evidence has consistently shown that peer ratings have high predictive validity (e.g., Kraut, 1975). Finally, some organizations also use persons outside the immediate work environment to observe individuals and then rate their performance. These sources include: (a) assessors in an assessment center, (b) field reviews conducted by people from a human resource department, and (c) evaluations from trainers (Latham & Wexley, 1981). One potential advantage of the use of outside observers is that it may reduce the randomness in evaluations that is due to appraisers use of different standards in evaluating performance.

In addition to the advantages that self-, peer, supervisor, and

observer ratings can provide individually as noted above, several potential advantages exist for their collective use in the performance appraisal process. Kane and Lawler (1978) advocate the use of multiple raters due to informational limitations, observational bias, and the non-randomness of performance sampled by the individual rater. In essence, content validity may be enhanced by tapping more of the behavioral domain of the job (Borman, 1974). Others have indicated that the use of mean ratings from multiple raters' scores would reduce halo or other measurement errors (Cooper, 1981; Miner, 1968). Further, the use of multiple rater sources may decrease subordinate defensiveness in performance appraisal interviews and increase accuracy in evaluations. Finally, interest and commitment may be enhanced because the use of multiple rater sources widens the participation of relevant persons in the performance appraisal process.

Research evidence supporting the use of these rater sources in the performance appraisal process is considerable. The use of supervisors as raters is clearly supportable on the basis of the necessity of supervisors to develop their subordinates, as well as to evaluate their progress (Latham & Wexley, 1981). Support for the use of peer ratings is also available in the literature (cf. Downey, Medland, & Yates, 1976; Fiske & Cox, 1960; Kaufman & Johnson, 1974; Kraut, 1975; Lewin & Zwany, 1976), while evidence for the usefulness of self-ratings appeared in a recent review by Mabe and West (1982). With respect to outside observers, Barrett (1966) concluded that evaluations done by outsiders can be based on a common frame of reference and are thus more likely than evaluations by supervisors to

be consistent across the organization.

Despite this evidence and the intuitive appeal of using multiple rater sources, research evidence directly comparing different rater sources has been inconsistent. Among those who have reported agreement in ratings of different rater sources are Holzbach (1978), Kavanaugh, MacKinney, and Wolins (1971), Mount (1984), and Williams and Seiler (1973). Differences in ratings of a group of raters have been reported by Borman (1974), Heneman (1974), Shore and Thornton (1986), and Thornton (1968). Other studies comparing multiple sources have found differences in discriminant validity, disagreement between factor structures for different rater sources, differences in rating strategies, and different degrees of halo and leniency (Baird, 1977; Bassett & Meyer, 1968; Blackburn & Clark, 1975; Borman, 1974; Griffiths, 1975; Holzbach, 1978; Ilgen, Peterson, Martin, & Boeschen, 1981; Kavanaugh, et al, 1971; Klimoski & London, 1974; Kraut, 1975; Meyer, 1980; Schneier & Beatty, 1978; Thornton, 1980; Tsui, 1983; Tsui & Ohlott, 1986; Wiley & Hahn, 1977; Williams & Seiler, 1973; Zammuto, London, & Rowland, 1982).

Many hypotheses have been advanced for the differences found among different rater sources. However, little noteworthy progress has been made in improving the quality of ratings across various rater sources. Further, despite suggestions advocating the use of ratings from several different rater sources, it is not clear why or when specific combinations of various rater sources should be effective (e.g., supervisor and peer ratings, supervisor and observer ratings, peer and self-ratings, peer and observer ratings). What is clear is that research results comparing ratings obtained from the various

rater sources are as yet too scarce and inconsistent to allow definitive conclusions regarding a best or most accurate rater source. In light of the importance of performance ratings to organizations, the current research attempts to isolate three factors (i.e., rater training, scale format, rating justification) which affect the psychometric properties of different rater sources. Specifically, the central concern of the present investigation centers around two questions: (1) What influence does rater training, scale format, and rating justification have on the quality of performance ratings (i.e., convergent validity, discriminant validity, halo, leniency) from different rater sources; and (2) Is there an interaction among these variables (i.e., rater training, scale format, rating justification) such that the quality of ratings from different rater sources can be enhanced by employing different combinations of these conditions?

This study continues research which spans over three decades. Therefore, it is important to review the research which has been completed in this area. Following a discussion of past findings, explanations for these results and the research hypotheses for the current study are presented.

Multiple Rater Source Research

A number of studies have investigated the quality of performance ratings given by different rater sources (see references cited on page 5). Most of these comparisons were made to address questions concerning the relative magnitude of psychometric properties (i.e., leniency, halo, range restriction, convergent and discriminant validity) attributable to ratings obtained from these rater sources. In essence, this research has examined the extent to which alternative

rater sources can agree with what has traditionally been used--supervisory ratings. The majority of this research has focused on self- and peer ratings. In fact, very little research could be found which used subordinates as a rater source (see Mount, 1984, Tsui, 1983, and Tsui & Ohlott, 1986 for exceptions), while no studies could be located which included outside observers as a rater source. Therefore, the findings regarding the two rater sources which have been investigated most frequently, peer and self-ratings, are reviewed below. The research findings are organized according to the most frequently examined psychometric properties.

Halo Effects. Halo has been conceptualized as a higher level of intercorrelation among rating dimensions than the true level of their intercorrelation (Saal, Downey, & Lahey, 1980). Halo, as a type of rater bias, occurs when a rater evaluates a person on all work dimensions using a global impression rather than specific examples corresponding to each dimension. The net result is that the person receives approximately equal scores on all dimensions. Ratings by supervisors consistently exhibit greater halo effects than self-ratings when the level of halo effect is measured by the magnitude of the intercorrelation among items obtained from each rater source (Heneman, 1974; Klimoski & London, 1974; Lawler, 1967; Tsui, 1983; Tsui & Ohlott, 1986). Peer ratings, on the other hand, tend to show comparable halo effects to supervisory ratings (Dickinson & Tice, 1973; Lawler, 1967) although Klimoski and London (1974) found a greater halo effect for peer ratings than for supervisory ratings.

Leniency Effects. Leniency error is a tendency to assign a higher rating to an individual than is warranted by the behavior of

that individual. Within the context of rater source research, the critical question concerning leniency centers around the extent to which one rater source provides higher ratings on a set of performance dimensions than other rater sources. A review by Thornton (1980) revealed that the preponderance of studies showed that individuals rate themselves higher than they are rated by other sources. Thornton's (1980) review indicated that these findings hold for several types of employees including clerical workers (Parker, Taylor, Barrett, & Martens, 1959), assemblers (Shore & Thornton, 1986), nurses (Klimoski & London, 1974), supervisors (Holzbach, 1978; Tsui, 1983; Tsui & Ohlott, 1986), and executives (Thornton, 1968). While there is some evidence that peer ratings are more lenient than supervisory ratings (e.g., Schneier, 1977), other research indicates that supervisor and peer ratings do not differ appreciably (e.g., Holzbach, 1978; Klimoski & London, 1974).

Convergent and Discriminant Validity. Convergent validity is defined as the extent of agreement between two or more measures of the same trait with different rating methods. Discriminant validity is defined as the extent of independent information provided by measures of different traits. In the context of rater source research the methods are defined by the rater sources, and the traits are defined by the dimensions on the rating instrument. There is some evidence that supervisor and peer ratings have reasonably high convergent and discriminant validity (Holzbach, 1978; Kavanaugh et al, 1971; Lawler, 1967). However, Borman (1974) and Zedeck, Imparto, Krausz, and Oleno (1974) found more disagreement than agreement between supervisor and peer ratings. The findings regarding convergent and discriminant

validity for self-ratings are also inconsistent. While Williams and Seiler (1973) found favorable convergent and discriminant validity for self- and supervisor ratings, Lawler (1967) and Nealy and Owen (1970) found little convergent or discriminant validity for self- and supervisory ratings.

In addition to the research reported above, Mount (1984) reviewed seven multiple rater source studies which used the multitrait-multimethod analysis of variance procedure (MTMM ANOVA) proposed by Kavanaugh et al. (1971) to assess the quality of performance ratings across rater sources. He found the median convergent validity to be .44, the median discriminant validity to be .17, and the median halo effect to be .47. These results however, are collapsed across rater sources which renders conclusions about individual sources impossible. More recently, a meta-analysis of MTMM studies of work performance ratings (Dickinson, Hassett, & Tannenbaum, 1986) exhibited mixed findings with respect to rater sources. Peer, self-, and subordinate ratings were associated with lower convergent validity, self-ratings were related to greater halo, and subordinate ratings had low discriminant validity.

To summarize, most of the available research indicates that evaluations given by various rater sources diverge. Supervisory ratings tend to be less lenient and contain more halo than either self- or peer ratings. The research evidence concerning convergent and discriminant validity is mixed. Finally, the inconsistent research findings in this area prevent conclusions about the superiority of any one type of rater source from being made at this time.

Factors Affecting Rater Source Agreement

A number of researchers have speculated on the reasons behind the inconsistencies among self-, peer, and supervisor ratings of performance. The view discussed most often is that disagreements stem from a tendency of different types of raters to base their ratings on different aspects of job performance or to weight factors of job performance differently (Klimoski & London, 1974; Latham & Wexley, 1981; Lawler, 1967; Mount, 1984; Tsui, 1983; Tsui & Ohlott, 1986). That is, "each rater occupies a different vantage point vis-a-vis the ratee" (Zammuto et al., 1982, p. 645).

Similarly, Guion (1965) suggested that raters in different positions may in fact be using different percepts or dimensions in their evaluations of an individual. This view is supported by Kavanaugh, Borman, Hedge, & Gould (1986) who posited that each rater source measures a part of the criterion space with more accuracy than the other rater sources and that no one position or organizational vantage point can provide the information necessary to determine a person's effectiveness. For example, self-ratings may be quite accurate for assessing job-relevant technical skills while supervisors may be best qualified to weigh an individual's performance across the various parts of the criterion space to reach an overall judgment (Kavanaugh et al., 1986). If incumbents, their peers, their supervisors, or outside observers observe work performance under different circumstances or even perceive the same performance differently, their separate perceptions of the individual's performance provide unique information. Collecting performance ratings from different rater sources, therefore, should increase the

amount of true performance variation that is measured.

An alternative view of this problem attributes the disparities in rating to systematic rater error (Holzbach, 1978; Saal, et al., 1980). It is hypothesized that certain rater sources may be more susceptible to some types of errors (e.g., leniency, halo, range restriction) than others. For example, self-ratings have frequently been found to contain less total variance than supervisory ratings (Thornton, 1980). Further, peer ratings have been found to be more lenient than supervisory ratings (Schneier, 1977). Finally, self-ratings have been found to contain less halo than either supervisory or peer ratings (Heneman, 1974; Klimoski & London, 1974; Tsui, 1983; Tsui & Ohlott, 1986). This last finding may help explain why ratings by several supervisors agree more than do self- and supervisor ratings. If supervisors' global assessments of an incumbent agree with one another, and these global assessments dominate their evaluations of the incumbent on specific performance dimensions, it is likely that supervisors will tend to agree with each other (converge) on each performance dimension. On the other hand, since incumbent ratings are more discriminating across dimensions, there is less of a global impression dominating their evaluations. Therefore, they are less likely to exhibit high agreement with supervisory ratings.

Another explanation for differences among supervisor, peer, and self-ratings is that the differences are caused by variant use of performance appraisal scales (Zammuto et al., 1982). That is, raters in different positions may erroneously conclude that different aspects of performance are relevant, or they may use the performance appraisal scales differently in rating performance. In a similar vein, it is

possible that the performance dimensions used in past research studies were not meaningful to different rater sources. Unfortunately, there is no way of directly assessing this latter possibility. Although some studies have had raters participate in dimension development, rarely have subordinates (who provide the self-ratings) been included in this process (see Dickinson & Tice, 1973 for an exception).

Alternative Explanations for Rater Source Differences

Each of the explanations provided above is a plausible argument for why differences exist among rater sources. However, little empirical evidence is available to support these views. Perhaps an altogether different approach is needed. Three issues were identified from the literature that form the basis for the current investigation.

1. Rater Training. The problems associated with rating errors (e.g., halo, leniency) have led researchers to call for the development of rater training programs to improve the quality of performance evaluations (e.g., Borman, 1979; DeCotiis & Petit, 1978; Dickinson et al., 1986; Dunnette & Borman, 1979; Kavanaugh et al., 1986; Smith, 1986). Rater training has recently shown some promise in improving the effectiveness of performance ratings (e.g., Borman, 1979; Bernardin & Pence, 1980; Dickinson & Silverhart, 1986; Fay & Latham, 1982; McIntyre et al., 1984; Pulakos, 1984). In fact, Kavanaugh et al. (1986) concluded that "it seems clear that it is not necessary to conduct research to determine if rater training should be a part of a performance measurement system. There must be some type of training..." (p. 36, underlining original). While research has now shifted to identify which types of training (e.g., psychometric error, accuracy) and methods of training (e.g., lecture, discussion) are most

effective, no research was found in the literature that attempted to assess the impact of rater training on the psychometric relationships among different rater sources.

Many of the arguments cited in the previous section for why differences exist among rater sources could be alleviated with rater training. It is unlikely that incumbents, peers, and supervisors have the same understanding of the overall goals and responsibilities of the individual being rated. In order to provide accurate ratings, different rater sources must be able to recognize examples of effective and ineffective performance, a goal that may be accomplished through rater training. Several authors (Bernardin & Buckley, 1981; Borman, 1979; Heneman, 1980) have suggested that possession of a common basis for rating may moderate rater agreement. One type of training, frame-of-reference (FOR) training (Bernardin & Buckley, 1981) is designed to "tune raters" to a common frame of reference so that worker behaviors can be similarly assessed by different raters. Bernardin (1981) found that FOR training actually increased interrater agreement, presumably by providing raters with a common basis for rating performance. As already noted, certain rater sources have been found to be more susceptible to some types of rating errors than others. If rating errors reduce total rating variance, then they directly restrict the covariance between two sources to reduce the agreement between the two sources of ratings (Mount, 1984). It seems reasonable to hypothesize, therefore, that convergence among sources would be enhanced by providing frame-of-reference training to all rater sources who will rate performance.

Many training programs to date have been successful in reducing

rating errors such as halo and leniency (e.g., Bernardin, 1978; Bernardin & Pence, 1980; Borman, 1975; Fay & Latham, 1982; Hedge, 1982; Ivancevich, 1979; Latham, Wexley, & Pursell, 1975; McIntyre et al., 1984; Pulakos, 1984). Smith (1986) found that 15 of the 19 studies he reviewed decreased halo with rater training. The most effective method for reducing halo was to include rater error training in the training program; while performance standards training was found to successfully reduce leniency in ratings (Ivancevich, 1979; Pulakos, 1984; Pursell, Dossett, & Latham, 1980). By providing training to rater sources, one would expect rating errors to be reduced, yet no research has examined this possibility.

To summarize, a number of studies have demonstrated that rater training programs can improve the effectiveness of at least some aspects of the performance rating process. It seems plausible to assume that the absence of a shared frame-of-reference would tend to exaggerate discrepancies between evaluators from different vantage points, since each must then supply his or her own frame-of-reference. As noted, rater training is ideally suited for developing a common frame-of-reference in performance evaluations and should improve the quality of ratings from all rater sources. Unfortunately, no research to date has examined the impact of rater training on the psychometric properties of different rater sources. The present study attempts to accomplish this goal. It is hypothesized that the quality of ratings (i.e., convergent validity, discriminant validity, halo, leniency) from the different rater sources will be enhanced when raters receive rater training.

2. Rating Formats. Numerous types of rating formats have been

developed in attempts to evaluate ratee performance accurately, alleviate the judgmental and measurement difficulties associated with performance appraisals, assist in providing feedback to ratees, and lessen the administrative burdens appraisals place on raters (Bernardin & Beatty, 1984). Formats aid in actual appraisals by determining the type and number of dimensions assessed, the types of judgments made, appraisal length, and comprehensiveness (Banks & Murphy, 1985). Graphic rating scales, checklists, forced-choice forms, forced-distribution forms, behaviorally anchored rating scales (BARS), and behavior observation scales (BOS) are some examples of the variety of methods psychologists have used to elicit performance ratings.

Comparisons of the psychometric properties of these different rating formats have resulted in inconclusive findings as to format superiority. In a narrative review of BARS, Kingstrom and Bass (1981) concluded that there was little difference between behavioral anchored scales and other formats. However, a recent meta-analysis of work performance ratings by Dickinson et al. (1986) yielded conclusions quite different from those reported by Kingstrom and Bass (1981). They found clear evidence that BARS and mixed standard scale (MSS) formats yielded higher quality ratings (i.e., greater convergent validity and/or lower method bias) than the graphic rating format. In addition, the use of behavioral dimensions was associated with higher convergent validity and lower method bias. Finally, the authors found that discriminant validity increased and method bias decreased as the number of ratings per dimension became greater.

These findings have important ramifications for research

attempting to explain discrepancies in the ratings of different rater sources. A review of the research revealed that most studies examining the psychometric properties of different rater sources have used some type of graphic rating scale (e.g., Heneman, 1974; Holzbach, 1978; Klimoski & London, 1974; Schneier & Beatty, 1978; Tsui, 1983; Tsui & Ohlott, 1986). Generalizing the Dickinson et al. (1986) findings, one might suggest that this may be a cause for the inconsistent and weak findings among different rater sources. Only four studies (Dickinson & Tice, 1973; Mascitti, 1978; Saal & Landy, 1977; Zedeck & Baker, 1972) could be found that used behaviorally-based rating scales (e.g., BARS, MSS, checklists) in assessing rater source errors. Although some research (e.g., Heneman, 1974; Klimoski & London, 1978; Mount, 1984) incorporated behavioral items, they were often confounded by the assessment of traits as well.

Both Dickinson and Tice (1973) and Zedeck and Baker (1972) examined the psychometric properties of different rater sources with behaviorally-based rating scales by means of the multitrait-multimethod (MTMM) approach. Participants in the Dickinson and Tice (1973) study were firefighters rated by supervisors and peers with a behavioral checklist; in Zedeck and Baker (1972), two nursing supervisory levels (head nurses and supervisors) evaluated staff registered nurses with a behavioral expectation scale (BES). An analysis of the multitrait-multimethod correlation matrix by means of the ANOVA approach (Kavanaugh et al., 1971) indicated that there was low convergent validity (ICC = .179) and low discriminant validity (ICC = .072) in the Dickinson and Tice (1973) data. Their results also indicated a moderate degree of halo (ICC = .273). Reanalysis of

the Zedeck and Baker (1972) data by Dickinson et al. (1986) yielded similar findings. The results revealed a high degree of convergent validity across supervisory levels (ICC = .396) and low discriminant validity (ICC = .075). In addition, there was moderate evidence for halo (ICC = .247).

Several possible explanations exist for these findings. First, neither study provided rater training suggesting that rater sources may have been employing different frames of reference. Both studies acknowledged this possibility. Also, Zedeck and Baker's (1972) use of supervisors as a rating source may not have been appropriate because a high percentage of a supervisor's time was spent in administrative functions, coordinating the activites in various areas of the hospital. Therefore, the supervisors did not have the same opportunity to observe and evaluate staff nurses as did the head nurses. This suggestion is supported by the fact that supervisors only contributed an average of 2.6 critical incidents per dimension in the development of the BES, whereas head nurses contributed an average of four incidents per dimension. Finally, the lack of discriminant validity in both studies may be explained by the considerable halo that existed.

Research by Mascitti (1978) and Saal and Landy (1977) represents the only studies found that assessed the impact of scale format on the rating errors of different rater sources. Mascitti (1978) obtained measures of job performance with a BARS and a numerically anchored rating scale (NARS) for self-, peer, and supervisory ratings. In comparing BARS and NARS for leniency, self- and peer ratings obtained on the NARS were more lenient than ratings on the BARS. In comparing

halo across rater sources, BARS and NARS produced different results. For BARS, immediate supervisors displayed less halo than the remaining sources, but there was no difference between secondary supervisory ratings and self-ratings, and no difference between self-ratings and peer ratings. For NARS, peer ratings showed greater halo than supervisory ratings, but no difference was found between supervisory and self-ratings.

Saal and Landy (1977), on the other hand, used police patrol officers to compare supervisory and peer ratings obtained by means of a mixed standard scale with ratings obtained on a behaviorally anchored scale. The criteria were leniency and halo. The mixed standard scale generally resulted in less leniency error and less halo error than the behaviorally anchored scale for both supervisor and peer ratings.

A major drawback to both studies was their sole use of bias (halo and leniency) as the criteria for determining format effectiveness. Cooper (1981) has explained that bias does not measure the effectiveness of a format as well as validity and accuracy. Also, various authors (e.g., Borman, 1979; Dickinson, 1986; Kavanaugh et al., 1986; McIntyre et al. 1984; Smith, 1986) have agreed that validity and accuracy are more appropriate criteria for evaluating format effectiveness. Consequently, neither the Mascitti (1978) nor the Saal and Landy (1977) study can make definitive conclusions with regards to the effects of scale format on the psychometric properties of rater sources.

Although the type of rating format has been a major topic of interest, Kavanaugh et al. (1986) recently concluded that the manner

in which the performance dimensions are described is also a critical feature of the performance rating instrument. In a review of some of the early research contrasting BARS with non-anchored graphic rating scales (e.g., Borman & Dunnette, 1975; Burnaska & Hollman, 1974; Campbell, Dunnette, Arvey, & Hellervik, 1973; Keaveny & McGann, 1975), Kavanaugh et al. (1986) made several suggestions: (1) the anchors or descriptors that define performance levels on job dimensions must be observable job behaviors or accomplishments; (2) these observables must be related to job-relevant tasks; and (3) the scale must be structured so that the rater can easily use it.

Quite clearly, research examining the psychometric properties of rater sources has not adhered to these suggestions. As already noted, most rater source research has employed a graphic rating scale. Consequently, the descriptors of different performance levels, if they in fact exist, are rarely observable job behaviors and are not related to job-relevant tasks. Further, a graphic rating scale does not provide for multiple ratings for each dimension, a recommendation made by Dickinson et al. (1986) for enhancing discriminant validity and reducing method bias.

To summarize, the predominant use of graphic rating scales in rater source research may help explain the poor agreement, low discriminant validity, and high rater bias typically found. Research that has used behavioral items (e.g., Mount, 1984) or behaviorally-based scales (e.g., Mascitti, 1978; Saal & Landy, 1977) has exhibited more promising results. Unfortunately, only a few of these studies exist. More research is needed to test the impact of scale format on the quality of ratings provided by different rater sources. Mascitti

(1978) and Saal and Landy (1977) provide the only research comparing rating formats. However, no research has assessed the impact of scale format on rater sources using validity as the criterion.

The current study was undertaken, in part, to compare the impact of two rating formats on different rater sources with validity as the criterion. Specifically, a traditional graphic rating scale was compared to a behavioral checklist (a description of both the checklist and the graphic rating scale will be provided later). It was hypothesized that the checklist would be superior to the graphic rating scale on all psychometric properties of interest. An examination of the characteristics of a behavioral checklist suggests several reasons for this prediction. First, among the recommendations posed by Dickinson et al. (1986) for improving convergent validity and reducing method bias were the use of behaviorally-oriented dimensions and behaviorally anchored scales. They also suggested the use of multiple ratings for each performance dimension to improve discriminant validity. The behavioral checklist used in the present study adheres to these recommendations while the graphic rating scale does not.

Further, Borman (1978) noted that "rating scale formats should conform to the cognitive processes raters utilize, and should not require raters to perform judgment steps they are incapable of making" (p. 143). Similarly, Smith and Kendall (1963) noted the necessity for all ratees to be evaluated in a comparable manner as well as for the necessity for raters to interpret the rating scales and their relationship to observable behavior in a similar fashion. It is believed that the use of a behavioral checklist will require the rater

to function less as a judge and more as an observer of behavior than a graphic rating scale. A behavioral checklist does not require as much information processing as a graphic rating scale. With a behavioral checklist the rater simply indicates the presence or absence of a number of behaviors, each associated with a specific dimension. Although the behavioral examples used in a checklist format are not necessarily identical to those a rater would observe, they serve as a concrete and specific frame of reference for the rater. On the other hand, with a graphic rating scale the focus is on the dimensions rather than the behaviors exibited. Here, the rater is forced to observe an episode of performance and infer from recalled behavior the performance of the ratee on several dimensions. Raters, in this instance, are left to form their own frame of reference. For these reasons, it is believed that the behavioral checklist will yield higher quality ratings (i.e., greater convergent and discriminant validity) than the graphic rating scale.

This study also hypothesizes an interaction between rater training and rating scale format. It is believed that the quality of performance ratings will be enhanced when raters receive training and use the behavioral checklist. The necessity of training raters to minimize rating errors and identify effective and ineffective behavior has already been noted. Once this has been accomplished, it is paramount that raters be provided with the tools (scales) that will allow them to use the skills that they have acquired.

3. Rating Justification. A number of studies have investigated the impact of the intended use of performance ratings on psychometric properties (e.g., McIntyre et al., 1984; Sharon & Bartlett, 1969;

Zedeck & Cascio, 1982). These studies have shown that ratings are more lenient under conditions of administrative use than under conditions of research use. Justification of a performance rating, on the other hand, is a variable which has received little attention in the performance appraisal literature. Wherry (1952) stated that "knowledge that the performance 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" (p. 13).

The majority of research in this area has been concerned with self-evaluations of performance although generalizations can be made to other rater sources. Mabe and West (1982) identified two measurement conditions frequently encountered in self-evaluation research that can be considered forms of justification: instructions of anonymity and expectation of validation. With respect to anonymity, self-enhancement motivation (the desire to enhance the perception of one's competence, Festinger, 1954) should be weaker when an individual's self-evaluation is anonymous than when the self-evaluation is not anonymous. An anonymous self-evaluation does not provide an external observer with specific information with which to judge the individual. It would therefore be expected that the individual has little reason to overestimate abilities, and more accurate self-evaluations should be given (Teachout, 1984). Similarly, the validity of self-evaluations could be improved by employing measurement conditions that include instructions that self-evaluations are to be compared with criterion measures (Mabe & West, 1982). In this instance, the incentive to report accurate self-evaluations would seem to be enhanced by the prospect that the self-

reports could be invalidated by comparison to other criterion measures (Mabe & West, 1982).

The belief that anonymous self-evaluations are more accurate has received some empirical support (Gordon & Petty, 1971; Sherwood, 1966; Sorenson, 1956; Teachout, 1984). Gordon and Petty (1971) found level of anonymity to significantly affect the accuracy of self-evaluations. Further, both Sherwood (1966) and Sorenson (1956) reported that anonymity improved the accuracy of self-evaluations by reducing the likelihood of socially desirable responses. These results suggest that anonymous responses are less inflated than identifiable responses. Apparently, individuals who could be identified were encouraged to self-enhance because they could benefit from a favorable self-evaluation.

On the other hand, not all research in this area has been supportive (Becker & Bakal, 1970; Sharon & Bartlett, 1969). Becker and Bakal (1970) used three sets of identification instructions on the MMPI lie scale and found that the anonymity instructions did not increase the prediction of distortion in responses. Sharon and Bartlett (1969) also found no differences in ratings of favorability between identified and unidentified individuals.

In addition to anonymity, it has also been hypothesized that an individual's self-evaluation is influenced by expectations that the self-evaluation will be subjected to validation. Evidence supporting this claim has been reported by several authors (Bassett & Meyer, 1968; Jones, 1973; Parker, et al., 1959; Regan, Gosselink, Hubsch, & Ulsh, 1975; Schlenker, 1975; Teachout, 1984). Schlenker (1975) found that self-evaluations were consistent with participants' expectations

of actual performance when objective events could invalidate an unrealistic positive self-evaluation. In addition, Teachout (1984) assessed the reading abilities of 120 undergraduate students who then made self-evaluations of their performance. He found that participants who expected their identifiable evaluations to be validated were more accurate. In contrast, when self-evaluations were not anonymous and validation was not expected, self-ratings of performance were more lenient (self-enhanced) and as such, were not accurate. It appears that the potential for objective validation tends to reduce the likelihood of self-enhancement and probably makes self-evaluations more realistic than those given in confidence.

In 1956 Dunnette and Heneman placed "justification" at one extreme of a continuum they labeled "psychological anonymity"; those who must justify their ratings are the least anonymous. However, most of the research that has been conducted in this area since then has focused on the anonymity end of the continuum and has not examined the influence of justification on the quality of performance ratings. Further, the focus of this line of research has clearly centered on self-evaluations of performance and has not examined the effects of justification on other rater sources.

The implications that "justification" may have on performance appraisal ratings for an organization are considerable. Performance appraisal feedback is often used by organizations to improve employee productivity and enhance development. However, feedback requires that the rater (in most cases the supervisor) justify his or her ratings to the incumbent. Stockford and Bissel (1949) found that supervisors who had to explain their ratings to their subordinates rated them more

leniently than when they did not have to explain them. If findings such as Stockford and Bissel's (1949) are true reflections of the impact that justification can have on performance ratings, one must suspect that ratings in organizations are inaccurate due to inflation caused by the influence of accountability. Consequently, the ability of an organization to differentiate among employees for promotion, training, and salary increases, is greatly hindered. Also, the use of inflated performance ratings in validation studies would adversely affect the results by reducing variability.

In addition to an examination of rater training and scale format, a third purpose of the current research is to examine the influence of justification on different rater sources. It was hypothesized that knowledge that raters would have to justify their ratings to the ratee would cause the rater to be more lenient than when the ratings would not have to be justified. However, one might expect rater training or scale format to work to offset the lenient ratings found when raters must justify their ratings. Consequently, this study also examined the interaction of rater training, scale format, and rating justification on the quality of performance ratings.

Research Hypotheses

Performance ratings by peers, incumbents, subordinates, and outside observers have been suggested as likely alternatives to the traditional supervisor-subordinate rating relationship. Research concerning the measurement of job performance by these different rater sources, however, is both scarce and inconsistent. As evidenced by the previous literature review, there are certain questions which still remain unanswered. Specifically, three areas of needed research

were identified (i.e., rater training, scale format, rating justification). Therefore, the current investigation attempted to assess the influence of rater training, scale format, and rating justification on the quality of performance ratings from different rater sources. Three rater sources were used: self, peer, and observer.

In addition, the research review noted several deficiencies in past studies, the most glaring of which was the use of bias as a criterion. The present study used a MTMM design which allowed for an examination of the construct validity of performance ratings by different rater sources. An analysis of rating data across rater sources on these indices, as well as leniency, provides useful information for evaluating the quality of the various sets of rating data. Such information, however, should not be directly interpreted to mean more or less accurate data from any specific source. Accuracy can be assessed only when a true performance score is available. The interest of this study is not accuracy per se, but the differential qualities of judgment made on the performance of ratees by different rater sources as affected by rater training, scale format, and rating justification.

In accordance with the objectives of this research, the following hypotheses were made:

1. Rater training will influence the leniency, halo, and convergent and discriminant validity shown by ratings from different rater sources. It is expected that self-, peer, and observer raters who receive rater training will exhibit less leniency and halo and more convergent and discriminant validity than those rater sources who

do not receive rater training.

2. Scale format will impact leniency, halo, and convergent and discriminant validity in a manner similar to rater training. Those rater sources who use the behavioral checklist will exhibit superior psychometric properties than rater sources who use the graphic rating scale.

3. The perception that performance ratings will have to be justified to the ratee will influence leniency, but no specific hypotheses are advanced with respect to halo or convergent and discriminant validity. Those sources who believe that they must justify their ratings will be more lenient than those raters who do not have to justify their ratings to the ratee.

4. Rater training and scale format will interact such that when self-, peer, and observer raters receive training and use the behavioral checklist, they will exhibit less leniency and halo and more convergent and discriminant validity than rater sources who do not receive rater training and use the graphic rating scale.

II. METHOD

Participants

Participants included 91 undergraduate students fulfilling a research requirement for a psychology course at Old Dominion University. Subjects also received $10 for their participation. These individuals took part in a role play exercise and later provided both self- and peer ratings of performance. These participants are referred to as ratees throughout the remainder of this study. Of the 91 ratees, 41% (37) were male and 59% (54) were female. Their ages ranged from 18 to 52 with a mean age of 24. Approximately 82% (75) were Caucasian, 11% (10) were Black, and 7% (6) were from other ethnic groups. Nine of the 91 ratees were freshman, 16 were sophomores, 37 were juniors, and 29 were seniors.

Participants also included 45 undergraduate psychology majors enrolled at the same university. They received extra course credit as well as $10 for participating in the study. These individuals served as the "observer" raters and did not take part in the role play exercise. Of the 45 observer raters, 45% (20) were male and 55% (25) were female. Their ages ranged from 19 to 36 with a mean age of 24. Approximately 93% (42) were Caucasian and 7% (3) were Black. Eleven of the observers were sophomores, 18 were juniors, and 16 were seniors.

Design

The design was a 2 x 2 x 2 x 3 x 4 fixed effects factorial with

two training conditions (training, no training), two scale formats (behavioral checklist, graphic rating scale), two levels of justification (justify, not justify), three rater sources (self, peer, observer) and four performance dimensions (problem analysis, problem solution, sensitivity, persuasiveness). Ratees were nested within training, scale format, and rating justification combinations.

Stimulus Exercise and Performance Dimensions

Stimulus Exercise Development.

A 10-minute interview simulation known as the customer role play served as the stimulus on which ratees were evaluated. This exercise was chosen because of its relevance to job situations in which individuals deal extensively with others on a one-to-one basis (Crooks, 1977). Support for the use of exercises of this nature can be found in Thornton and Byham (1982). These authors have estimated that over 75% of all assessment centers use an interview simulation similar to the customer role play. In addition, according to Thornton and Byham (1982) an interrater reliability coefficient of .80 has been reported for the interview simulation (Russell & Byham, 1980), and two unpublished studies were reported by Thornton and Byham (1982) to have strong correlations between the interview simulation and overall assessment center ratings.

In the exercise used in the present study, the ratee assumed the role of a store manager who had to solve the problem of an irate customer (see Appendix A for the role play instructions provided to each participant). The irate customer was played by a male graduate student enrolled in the Ph.D. program in industrial/organizational psychology. Prior to participating in the research study, each ratee was told that: (a) the study involved performance appraisals, (b) he

or she would be asked to participate in a role play exercise, and then, (c) return within three weeks to rate videotaped performances of both themselves and their peers (see Appendix B). All individuals who agreed to participate signed an informed consent form and then took part in the customer role play exercise. Upon completion of the exercise ratees were told when to return to provide performance ratings. In all, 96 videotapes were produced.

Performance Dimensions. Four performance dimensions were identified for use with the customer role play on the basis of past reviews of the assessment center literature (Dickinson & Silverhart, 1985; Thornton & Byham, 1982). The dimensions used for evaluation included: problem analysis, problem solution, sensitivity, and persuasiveness. Dimension definitions appear in Appendix C.

The identification of these dimensions was supported by the work of Thornton and Byham (1982) who reviewed over 1,000 assessment center reports in 12 large organizations. They found that in approximately 90% of the interview simulations conducted, the performance of the ratee on the dimensions of problem analysis, problem solution, sensitivity, and persuasiveness could be reliably evaluated by assessors.

Rating Scales

Two types of rating scales, a behavioral checklist and a traditional graphic rating scale, were used in the present study to measure the performance of ratees in the role play exercise. Raters used either the behavioral checklist or the graphic rating scale.

Behavioral Checklist. A modified version of a behavioral checklist developed by Campbell (1986) was used in the present study.

The development of the checklist occurred over a three-stage process. At each stage, 3 to 6 Old Dominion University graduate students familiar with the role play exercise participated in scale development. Each stage in the development of the behavioral checklist is briefly discussed below. For a detailed description, see Campbell (1986).

Stage 1. Critical incidents were generated by three of the graduate students who had viewed eight videotaped customer role plays obtained prior to Campbell's (1986) research. Incidents were then edited to remove redundancies. After the editing process, 219 behavioral items remained.

Stage 2. Six graduate students familiar with the role play exercise were then provided with a list of the dimensions chosen for inclusion in the study (i.e., problem analysis, problem solution, sensitivity, and persuasiveness). The six graduate students met to discuss the dimensions and identify key words that were used to convey information on the context in which a behavior was displayed. Following this meeting, they were asked to assign each behavioral item to the most representative dimension. A behavior was retained if 75% of the judges agreed on the assignment of the behavior to a dimension. One hundred and seven items were eliminated during this process.

The same group of judges was then asked to rank order, within dimensions, the remaining 112 behaviors from effective to ineffective. Agreement of the rankings was evaluated by means of Kendall's Coefficient of Concordance (W). As suggested by Taylor (1968), a reliability coefficient of .75 or greater was used to ensure unambiguous dimensions. Each of the four dimensions satisfied this

criterion.

Stage 3. Means and ranges were computed for each behavioral statement's rank. Items with ranges of 15 or less were considered for inclusion on the checklist. These 102 items were ranked within dimensions from lowest to highest and divided into five groups representing approximately equal intervals of effectiveness as measured by the mean ranks. Numerical weights of 1 to 5 were assigned to each item corresponding to its level of effectiveness, with one being the least effective and five the most effective behavior. For each dimension three items were selected from each level of effectiveness. Thus, each dimension consisted of 15 items. The behavioral items were placed under dimension headings in the order in which they were expected to occur to help aid the rater in evaluation. Four dimension scores were obtained for each ratee by calculating the arithmetic mean of the weights for each item checked in a dimension. The behavioral checklist appears in Appendix D.

Graphic Rating Scale. The alternative method against which the behavioral checklist was compared was a graphic rating scale. In a study comparing rating scale formats, Borman and Vallon (1974) concluded that formats that included both dimension definitions and verbal descriptions of the numbers on the scale were superior to formats that did not possess these characteristics. Therefore, the graphic rating scale used in this study contained a definition of each performance dimension as well as a verbal description of each number on the scale. All dimensions were rated on a five-point scale ranging from much less than acceptable (1) to much more than acceptable (5). The graphic rating scale appears in Appendix E.

Rater Training

Prior to rating the videotapes all raters reviewed the ratees' role in the exercise by reading the instructions presented to ratees before they participated in the role play exercise (see Appendix A). In addition, all raters received definitions of each dimension to be rated. After the raters had an understanding of the roles and dimensions involved in the exercise, the rating formats were introduced to the raters with instructions concerning their use. Raters using the behavioral checklist were first asked to take a few minutes to familiarize themselves with the behavioral items listed under the dimension headings. Raters utilizing the graphic rating scale, on the other hand, were first asked to familiarize themselves with the dimension definitions. These instructions, which appear in Appendix F, represented the only difference in format training. Following these instructions, raters in the no-training condition viewed and rated six videotapes.

The suggestions and results of recent research investigations guided the development of the rater training program in the present study (e.g., Bernardin, 1981; Dickinson et al., 1986; Kavanaugh et al., 1986; Latham & Wexley, 1981; McIntyre et al., 1984; Pulakos, 1984; Smith, 1986). Smith (1986) and Latham and Wexley (1981) suggested that if a training program is to bring about a permanent change in rater behavior, it must incorporate rater participation, feedback, and rating practice using the formats. Providing raters with the opportunity to participate in a group discussion along with practice and feedback produces better results than presenting the training material to raters through a lecture (Smith, 1986).

Consequently, detailed practice and feedback was provided to all raters in the training condition to aid the different rater sources in developing common standards of effective performance in the role play exercise. A description of the training program follows.

Following a brief introduction to the training session (Appendix G), raters were provided with a list of behaviors that are typically exhibited for the dimensions in the role play exercise. These behaviors were identical to those in the behavioral checklist. Raters were asked to take a few minutes to familiarize themselves with the behaviors under the dimension headings. All raters in the training condition, regardless of format used, then practiced by rating a videotape of a customer role play exercise one dimension at a time. Ratings were discussed as to what particular ratee behaviors led raters to their ratings. Any problems encountered were discussed and corrected at this time. If necessary, selected portions of the videotape were viewed again. Finally, raters were asked to rate a videotape on all dimensions.

This portion of the training program combined portions of Performance Dimension Training and Performance Standards Training (Smith, 1986). Performance Dimension Training attempts to improve the effectiveness of ratings by familiarizing raters with the dimensions by which performance is rated. This was accomplished through extensive practice and feedback on both the dimensions and the rating scales. Performance Standards Training attempts to provide raters with a frame of reference for making evaluations of the ratees' performance. In this study raters compared their practice ratings with ratings provided by the experimenter and the ratings of others in

their training group. Both methods have been found to improve the quality (i.e., reduce halo and leniency, improve accuracy) of performance ratings (e.g., Fay & Latham, 1982; McIntyre et al., 1984; Pulakos, 1984; Pursell et al., 1980). It was believed that this training component would provide different rater sources with a common frame of reference for considering ratee performance as well as a complete understanding of the performance dimensions to be rated. This training component lasted approximately 60 minutes.

Because raters in the justification condition expected to have to justify their performance ratings to their peers, "justification training" was provided as a final training component. This training included instructions to observe performance carefully, watch for specific behaviors, and to take notes. These instructions occurred during the introductory phase of rater training (see Appendix G). Following the practice and feedback portion of the training program a short lecture on effective feedback skills was also provided to aid raters in the group discussion of ratings that was to occur later in the study. Characteristics of effective feedback focused on in this lecture included: (1) the need to be specific, (2) the need to focus on behaviors rather than personality, and (3) the need to be prepared. Appendix H provides a script of this lecture. This training component lasted approximately 20 minutes.

Rating Justification

The justification condition was manipulated by reading one of two instructional sets to the raters. Instructions were the same for all rater sources. Raters received verbal instructions during the introductory phase of the experiment. The following instructions

produced the justification conditions:

1) No Justification. These ratings will be used for research purposes only. They will not be used to evaluate you or your peers in any way. Your ratings will be strictly anonymous. Do not place your name on the rating form or identify yourself in any manner. This study is part of a doctoral dissertation on the rating process being conducted by S. Woods of the Department of Psychology.

2) Justification. Write your name and social security number on the rating form in the space provided. Your ratings will be used in a feedback discussion among yourself, your peers, and the experimenter to help improve the ability of individuals in your group to rate performance effectively. Past experience has shown that face-to-face discussions are very successful for improving performance. You will therefore be required to justify your ratings in the group discussion.

To ensure that the justification condition was appropriately perceived, written instructions also appeared on the cover page of the rating form. Raters were asked to read these instructions prior to viewing the videotapes. All raters received a letter approximately three weeks after data collection was completed disclosing the full nature of the experiment.

Rating Procedure

Five of the 96 undergraduates who participated in the role play exercise failed to return to provide performance ratings. Hence, self-ratings were obtained from 91 ratees. Peer and observer ratings,

however, were obtained on all 96 role play participants.

Specifically, self- and peer ratings were obtained in 16 three and one-half hour sessions conducted by the same experimenter. Each session was divided into three phases. In the first phase, raters were reminded of the study's purpose. Raters were told that the study involved performance appraisals and that they would be rating videotaped performances of both themselves and their peers. In addition, raters received verbal instructions designed to manipulate their expectations that they would have to justify their performance ratings. These instructions were described above.

At the conclusion of this introductory phase a 30-minute small group exercise was held for all raters who had participated in the role play exercise. This group exercise was provided to allow group members to get acquainted quickly by sharing their initial concerns and expectations with one another. Specifically, the small group exercise provided the ratees with an opportunity to: (1) review the role play exercise; (2) assess and discuss their initial concerns, anxieties, and expectations regarding their participation in the role play; (3) list and discuss the difficulties they encountered during the role play; and (4) list and discuss the strategies/approaches they used in dealing with the irate customer. Throughout the small group exercise the group discussion focused on the common experience that they had all shared during the role play exercise. Appendix I outlines the procedure followed during the group exercise.

Phase two consisted of rater training. A description of the training program, which lasted approximately 60 minutes, appeared above. To briefly review, all raters received definitions of the

performance dimensions and instructions in the use of the rating form provided by the experimenter. Those raters in the training condition also received a short lecture on effective feedback skills as well as practice and feedback in rating performance. Finally, in phase three, raters provided performance ratings on the six videotaped customer role plays which corresponded to the members in their experimental condition.

"Observer" ratings were provided by 45 undergraduate students after the 96 role plays had been videotaped. These 45 raters were randomly assigned to one of the 16 experimental conditions. Thirteen of the experimental groups were comprised of three observer raters each. However, while three observer raters were assigned to the three remaining groups, three individuals failed to attend as scheduled. Consequently, these three experimental groups were comprised of two observer raters each. Rater training was identical to that provided to self- and peer raters. Upon the completion of rater training, each group of observer raters viewed six videotapes and provided ratings on the dimensions using the format provided. Justification was manipulated as described above.

Manipulation Checks

To assess rater comprehension of the training program, a two-part test (see Appendix J) was designed and administered to all raters before and after the viewing of the six experimental videotapes. Specifically, the pre-test was administered prior to the format instructions while the post-test was administered immediately following the completion of the rating session. Part I consisted of 30 items which asked the rater to match each performance dimension

with a behavioral component. The behavioral components were actual items from the behavioral checklist described above. These items were randomly assigned to the pre- and post-tests. A total score of 30 was possible for Part I.

Part II consisted of two open-ended questions designed to tap comprehension of the "justification training": (1) If you were responsible for observing and then rating an individual's performance, what are some of the things you would do to make sure your rating was accurate?; and (2) People often receive performance feedback from their supervisor in a formal performance appraisal feedback interview. What do you believe are some important components of an effective feedback discussion? Question 1 was worth three points while Question 2 was worth five points for a total possible score of 8 on Part II.

Mean scores on the pre- and post-tests for both the training and no training groups are presented in Table 1. An examination of Table 1 reveals no difference between training groups for the pre-test. However, as expected, the training program significantly affected comprehension of the training information for both Part I ($t(134) = 4.97$, $p < .01$) and Part II ($t(134) = 14.93$, $p < .01$) of the post-test, as well as the combined score ($t(134) = 10.74$, $p < .01$). This finding confirms that the training program was effective in communicating the training information to raters in the training groups.

In addition to the training test, a 19-item post-experimental questionnaire was administered to all raters upon the completion of the post-test (see Appendix K). Ten items assessed the effectiveness of the experimental manipulations (i.e., training, scale format, rating justification), five items assessed demographic information,

Table 1

Training vs No Training Comparison of the Pre- and Post-Tests.

	Means		t-value
	Training	No Training	
Pre-test			
Part I	21.46	21.99	.76
Part II	.14	.09	1.00
Combined	21.61	22.07	.68
Post-test			
Part I	25.23	22.49	4.97*
Part II	3.95	.19	14.93*
Combined	29.23	22.69	10.74*

Note. Degrees of freedom for all t-tests were 134.

*$p < .01$.

and four items were added for face validity.

Mean scores for the items of interest are presented in Table 2. Questions 7, 11, 12, 16, and 18 were designed as manipulation checks for the justification condition. As expected, mean scores for raters in the justification condition were significantly higher than the scores of raters who did not have to justify their ratings on Questions 7, 11, 12, and 16. In addition, when asked what their ratings would be used for (Question 18), all raters responded appropriately. That is, raters in the no justification condition believed that their ratings would be used for psychological research, while those in the justification condition believed that their ratings would be used in a feedback discussion group. Questions 8 and 10 were designed to check for differences in rating confidence due to training. On question 8, raters who received training were more confident in assessing an individual's performance than raters who did not receive training ($t(134) = 9.59$, $p < .01$). However, there was no difference between training groups on Question 10 ($t(134) = .53$). Finally, Questions 13, 15, and 17 assessed scale format/instruction adequacy. Although there was no difference between format groups for Question 15, as indicated in Table 2, raters who used the behavioral checklist believed that they were better able to document an individual's performance than raters who used the graphic rating scale ($t(134) = 11.81$, $p < .01$). Also, both format groups felt that instructions for the rating formats were clear and easy to understand (Question 17). These results, combined with the training test results, suggest that the experimental manipulations in this study were successful.

Table 2

Analyses Summarizing Results of the Post-Experimental Questionnaire.

	Means		
	Justif Condition	No Justif Condition	t-value
Question 7 (can your ratings be matched with your name)	4.38	1.33	22.32*
Question 11 (will you be held accountable)	.94	.10	17.84*
Question 12 (can you be identified)	3.88	1.51	13.24*
Question 16 (will your peers know what ratings you gave)	.62	.06	8.51*
	Train	No Train	t-value
Question 8 (how confident were you)	3.73	2.24	9.59*
Question 10 (how confident that your ratings were accurate)	2.93	2.85	.53
	Behavioral Checklist	Graphic Scale	t-value
Question 13 (could you adequately document performance)	3.81	1.97	11.81*
Question 15 (how confident that your ratings were accurate)	2.88	2.76	.78
Question 17 (were the instructions for format use clear)	4.28	4.13	.90

Note. Abbreviations are: Justif = Justification; No Justif = No justification; Train = Training; No Train = No training. Degrees of freedom for all t-tests were 134.

*$p < .01$.

III. RESULTS

Analytic Approach

Multitrait-Multirater Analyses. The primary objective of the present study was to examine the influence of rater training, scale format, and rating justification on the quality of performance ratings provided by different rater sources (i.e., self, peer, observer). Convergent validity, discriminant validity, and halo effects were tested with the multitrait-multimethod analysis of variance (MTMM ANOVA) approach proposed by Kavanaugh et al. (1971). This approach was selected in favor of the correlational approach advocated by Campbell and Fiske (1959) because it provides a more efficient method of summarizing and interpreting the evidence for construct validity. Since there were multiple peer and observer ratings on each performance dimension for a particular ratee, the responses were summed within rater source and the arithmetic mean was calculated for each dimension. These values represented the peer and observer ratings for each ratee.

A number of studies have used the MTMM ANOVA procedure to analyze the construct validity of different rater sources (e.g., Heneman, 1974; Holzbach, 1978; Mount; 1984). In this instance, the multimethods are defined by the rater sources, and the multitraits are defined by the dimensions on the rating instrument (Note: When raters are used as methods in the MTMM design it is abbreviated MTMR). Convergent validity reflects agreement among rater sources in assessing dimensions of behavior. Discriminant validity reflects the

differential ordering of ratees by dimensions. Method bias (halo) indicates a differential ordering of ratees by rater sources. Most studies have found evidence for convergent validity, very little support for discriminant validity, and a large halo effect.

To assess the influence of training, format, and justification on the quality of performance ratings exhibited by the different rater sources, each experimental condition was treated as a mini-MTMR design and separate ANOVAs were performed. Table 3 presents a summary of the MTMR design including a psychometric interpretation for each source of variation. Of particular interest are the random effects of Ratees (convergent validity), Ratees x Dimensions (discriminant validity), Ratees x Rater Source (halo effect), and Error. These sources provide information about the validity of the measures and allow inferences about individual differences among ratees. In all, eight ANOVAs were performed.

In addition, variance components and intraclass correlation coefficients (ICC) were computed for each source of variance within an experimental condition. Variance components are computed in order to make inferences about the magnitude of the effects obtained in the analysis of variance. They provide a comparison of the relative sizes of convergent validity, discriminant validity, halo, and error while controlling for degrees of freedom. The computation of ICCs, on the other hand, permits comparisons across experimental conditions. Each ICC estimates the proportion of variance accounted for by that source relative to the variation accounted for by all sources (Dickinson, 1986). The variance components were computed according to the procedures set forth by Vaughan and Corballis (1969); while the ICCs

Table 3

Summary Table and the Psychometric Interpretations of the MTMR Design Within Each Experimental Condition.

Source	Psychometric Interpretation
Dimensions (D)	Dimension Bias
Rater Source (S)	Source Bias
D x S	Dimension by Source Bias
Ratees (R)	Convergent Validity
R x D	Discriminant Validity
R x S	Halo Effect
Error	Sampling and Measurement Errors

were calculated as the ratio of a source's variance component to the sum of all relevant variance components (Bartko, 1966).

Leniency Analyses. Several conceptualizations of leniency exist (Saal et al., 1980). However, the notion that ratings are consistently too high pervades most of these conceptualizations. In the present study, leniency was operationally defined as the extent to which one rater source provides higher ratings on a set of performance dimensions than the other rater sources.

A 2 (Training) x 2 (Format) x 2 (Rating Justification) x 3 (Rater Source) x 4 (Dimensions) ANOVA was computed to test the effects of the experimental conditions on the leniency of performance ratings across rater sources. Hypotheses 1 through 4 proposed that differences in the level of ratings across rater sources would be influenced by training, scale format, rating justification, and the training by format interaction. Of concern in this analysis is the rater source effect (S) and its interaction with other sources of variation (e.g., Training x Source, Format x Source, Justification x Source, Training x Format x Source). Simple effects tests and Tukey (hsd) post hoc tests were computed where appropriate. Table 4 provides a summary of this design describing the sources of variation and their error term.

The General Linear Model program in SAS (SAS User's Guide, 1985) was used for the leniency analyses. This program, and others like it, can not handle an unbalanced design as large as the one used for the present study. Consequently, the five missing self-ratings on the four performance dimensions were replaced with the appropriate cell means. This replacement procedure accounted for less than 2% of the total observations in the study.

Table 4

Analysis of Variance Summary Table for the Design Used to Test for Leniency Effects.

Source	Error Term
Training (T)	R/TxFxJ
Format (F)	R/TxFxJ
Justification (J)	R/TxFxJ
Rater Source (S)	S x R/TxFxJ
Dimensions (D)	D x R/TxFxA
Ratees (R)/TxFxJ	No Term
T x F	R/TxFxJ
T x J	R/TxFxJ
T x S	S x R/TxFxJ
T x D	D x R/TxFxJ
F x J	R/TxFxJ
F x S	S x R/TxFxJ
F x D	D x R/TxFxJ
J x S	S X R/TxFxJ
J x D	D x R/TxFxJ
S x D	S x D x R/TxFxJ
T x F x J	R/TxFxJ
T x F x S	S x R/TxFxJ
T x F x D	D x R/TxFxJ
T x J x S	S x R/TxFxJ
T x J x D	D x R/TxFxJ

Table 4 (Concluded)

Source	Error Term
T x S x D	S x D x R/TxFxJ
F x J x S	S x R/TxFxJ
F x J x D	D x R/TxFxJ
F x S x D	S x D x R/TxFxJ
J x S x D	S x D x R/TxFxJ
T x F x J x S	S x R/TxFxJ
T x F x J x D	D x R/TxFxJ
T x F x S x D	S x D x R/TxFxJ
T x J x S x D	S x D x R/TxFxJ
F x J x S x D	S x D x R/TxFxJ
T x F x J x S x D	S x D x R/TxFxJ
D x R/TxFxJ	No Term
S x R/TxFxJ	No Term
S x D x R/TxFxJ	No Term

MTMR Results

As noted above, the ANOVA technique described by Kavanaugh et al. (1971) was used to quantify the relative contribution of the various sources of variance for each of the experimental conditions. Eight ANOVAs were computed in the present study and are summarized in Tables 5 through 12. Of interest are the random effects of Ratees (convergent validity), Ratees x Dimensions (discriminant validity), and Ratees x Rater Source (halo). The tables indicate that the Ratees and Ratees x Dimensions sources of variation were highly significant in all eight experimental conditions ($p < .01$). There is differentiation among ratees attributable to the rater sources, that is, person variance or convergent validity. The Ratees x Dimensions interaction indicates a differential ordering of the ratees on the four performance dimensions. Thus, there is evidence for discriminant validity.

Finally, Tables 5 to 12 reveal a significant Ratees x Rater Source effect (halo) in all four experimental conditions where the graphic rating scale was used. In contrast, there was no evidence for halo in any of the experimental conditions where the behavioral checklist was used. The significant halo effect in those circumstances where the graphic rating scale was used indicates that ratees were ordered differently by different rater sources. This finding confounds interpretation of the Ratees effects (convergent validity). That is, the differential ordering of the ratees may be due to "halo" errors committed by some of the rater sources rather than overall differences across dimensions.

Hypotheses 1, 2, and 4 were concerned with the influence of rater

Table 5

Summary Table for the MTMR Analysis of Performance Ratings for the No Training-Graphic Rating Scale-No Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	1.995	3.62*	.033	.050
Rater Source (S)	2	.520	0.65	-.004	.000
D x S	6	.487	3.08*	.015	.023
Ratees (R)	10	2.085	13.17**	.161	.246
R x D	30	.551	3.48**	.131	.200
R x S	20	.802	5.07**	.161	.246
Error	60	.158		.158	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the sources coefficient was set to zero. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

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

Table 6

Summary Table for the MTMR Analysis of Performance Ratings for the No Training-Graphic Rating Scale-Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	5.587	8.41**	.103	.109
Rater Source (S)	2	12.251	26.83**	.165	.174
D x S	6	.564	2.51*	.014	.015
Ratees (R)	11	3.124	13.89**	.242	.256
R x D	33	.665	2.95**	.147	.155
R x S	22	.457	2.03*	.050	.053
Error	66	.225		.225	

Note. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

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

Table 7

Summary Table for the MTMR Analysis of Performance Ratings for the No Training-Behavioral Checklist-No Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	2.642	2.76	.042	.059
Rater Source (S)	2	.913	2.95	.010	.014
D x S	6	.300	1.06	.001	.001
Ratees (R)	9	1.943	6.84*	.138	.195
R x D	27	.956	3.36*	.224	.317
R x S	18	.310	1.09	.006	.009
Error	54	.284		.284	

Note. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

*p < .01.

Table 8

Summary Table for the MTMR Analysis of Performance Ratings for the No Training-Behavioral Checklist-Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	4.872	7.45*	.088	.153
Rater Source (S)	2	.287	1.39	.001	.001
D x S	6	.098	.49	-.004	.000
Ratees (R)	11	1.812	9.16*	.134	.235
R x D	33	.654	3.31*	.152	.266
R x S	22	.206	1.04	.002	.003
Error	66	.198		.198	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the sources coefficient was set to zero. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

*$p < .01$.

Table 9

Summary Table for the MTMR Analysis of Performance Ratings for the Training-Graphic Rating Scale-No Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	1.672	2.35	.020	.030
Rater Source (S)	2	1.047	1.81	.007	.011
D x S	6	.179	.71	-.003	.000
Ratees (R)	11	1.982	7.88**	.144	.220
R x D	33	.712	2.83**	.154	.235
R x S	22	.579	2.30*	.082	.125
Error	66	.252		.252	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the sources coefficient was set to zero. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

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

Table 10

Summary Table for the MTMR Analysis of Performance Ratings for the Training-Graphic Rating Scale-Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	1.138	1.70	.011	.014
Rater Source (S)	2	.093	.14	-.008	.000
D x S	6	.109	.37	-.008	.000
Ratees (R)	10	3.398	11.66**	.259	.340
R x D	30	.669	2.30*	.126	.166
R x S	20	.652	2.24*	.090	.118
Error	60	.291		.291	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the sources coefficient was set to zero. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

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

Table 11

Summary Table for the MTMR Analysis of Performance Ratings for the Training-Behavioral Checklist-No Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	.698	.63	-.010	.000
Rater Source (S)	2	.243	1.71	.002	.004
D x S	6	.108	.83	-.001	.000
Ratees (R)	10	1.346	10.34*	.101	.182
R x D	30	1.116	8.57*	.329	.594
R x S	20	.142	1.09	.003	.005
Error	60	.130		.130	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the sources coefficient was set to zero. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

*$p < .01$.

Table 12

Summary Table for the MTMR Analysis of Performance Ratings for the Training-Behavioral Checklist-Justification Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	3	.819	.58	-.013	.000
Rater Source (S)	2	.498	.98	-.000	.000
D x S	6	.521	1.75	.009	.010
Ratees (R)	11	2.877	9.64*	.215	.229
R x D	33	1.423	4.77*	.375	.400
R x S	22	.508	1.70	.052	.055
Error	66	.299		.299	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the sources coefficient was set to zero. Abbreviations are: df = degrees of freedom; MS = Mean squares; VC = Variance component; ICC = Intraclass correlation coefficient.

*$p < .01$.

training, scale format, and their interaction on the quality of performance ratings given by the three rater sources. Consequently, the relative amount of variation accounted for by each experimental effect was evaluated by comparing variance components and intraclass correlation coefficients (ICCs) as noted above. Dickinson et al. (1986) suggested the following verbal description when interpreting intraclass correlation coefficients: high, good (above .30), medium, moderate (.20 to .29), and low, poor (less than .20). Table 13 presents the ICC values for convergent validity, discriminant validity, and halo for each of the eight experimental conditions.

Hypotheses 1, 2, and 4 can be evaluated by examining the ICC values in Table 13. With respect to Hypothesis 1, it was suggested that the ICC values for convergent and discriminant validity would be higher in those instances where raters received training than in those circumstances when no training was provided. Further, it was hypothesized that the ICC value for halo would be lower for those raters who received training. An examination of Table 13 reveals some support for this hypothesis. The ICC values for discriminant validity were generally higher for those who received training (M ICC = .349) than for those rater sources who did not receive training (M ICC = .235). The difference in discriminant validity can be attributed to the high ICC values in the two training conditions that used the behavioral checklist (ICC for Training-Behavioral Checklist-Justification = .400; ICC for Training-Behavioral Checklist-No Justification = .594). Contrary to Hypothesis 1, however, there was no difference in convergent validity and halo between training conditions. Mean ICC values for training and no training groups were

Table 13

Comparison of ICC Values for Convergent Validity, Discriminant Validity, and Halo Across Experimental Conditions.

	Training GRS Justif	Training GRS No Justif	Training BC Justif	Training BC No Justif
Convergent Validity	.340	.220	.229	.182
Discriminant Validity	.166	.235	.400	.594
Halo Effect	.118	.125	.055	.005

	No Train GRS Justif	No Train GRS No Justif	No Train BC Justif	No Train BC No Justif
Convergent Validity	.256	.246	.235	.195
Discriminant Validity	.155	.200	.266	.317
Halo Effect	.053	.246	.003	.009

Note. Abbreviations are: GRS = Graphic rating scale; BC = Behavioral checklist; Justif = Justification; No Justif = No justification; No Train = No training.

essentially equal in magnitude for both convergent validity and halo (M ICCs = .243 and .076 compared to .233 and .078, respectively).

Hypothesis 2 was concerned with the influence of scale format on the psychometric properties of ratings from different rater sources. Specifically, it was believed that those rater sources who used the behavioral checklist would exhibit greater convergent and discriminant validity and lower halo than rater sources who used the graphic rating scale. Table 13 indicates some support for this hypothesis. As anticipated, discriminant validity was high in those situations where the behavioral checklist was used (M ICC = .394) and low in those situations where the graphic rating scale was employed (M ICC = .189). In addition, ICC values for halo were generally higher among graphic rating scale users when compared to raters who used the behavioral checklist (M ICCs = .136 and .018, respectively). Contrary to Hypothesis 2, however, the ICC values for convergent validity were generally higher for those who used the graphic rating scale (M ICC graphic scale = .266; M ICC checklist = .210).

Hypothesis 4 suggested that rater training and scale format would interact such that when self-, peer, and observer raters received training and used the behavioral checklist, they would exhibit less halo and more convergent and discriminant validity than rater sources who did not receive training and who used the graphic rating scale. Once again, mixed support was found. As hypothesized, discriminant validity was highest in the Training-Behavioral Checklist conditions (M ICC = .497). The next highest ICC values were for those raters in the No Training-Behavioral Checklist condition (M ICC = .292). Further, discriminant validity was higher when training was provided

and the graphic rating scale was used (M ICC = .201) than when no training was given and the graphic rating scale was used (M ICC = .178).

With respect to halo and Hypothesis 4, it should be noted that in seven of the eight experimental conditions there was little evidence of halo (ICCs less than .20). The exception was the No Training-Graphic Rating Scale-No Justification condition which had a moderate halo effect (ICC = .246). Specifically, the ICC values for halo were generally lower in those instances where training was received and the behavioral checklist was used (M ICC = .030) than when the graphic rating scale was used, regardless of whether or not training was provided (M ICC Training-Graphic Rating Scale = .122; M ICC No Training-Graphic Rating Scale = .150). Somewhat unexpectedly, the lowest degrees of halo were found in the no training conditions. However, the use of the behavioral checklist in both instances probably worked to offset the lack of training. Finally, contrary to Hypothesis 4, the lowest ICC values for convergent validity were found among those who received training and used the behavioral checklist (M ICC = .206). The largest ICC values for convergent validity, on the other hand, were found among those who used the graphic rating scale (M ICC Training-Graphic Rating Scale = .280; M ICC No Training-Graphic Rating Scale = .251). Apparently, use of the behavioral checklist resulted in somewhat lower convergent validity even when training was provided.

Finally, Table 13 provides some insights into the effect that rating justification had on the quality of ratings provided by different rater sources. Although no specific hypotheses were

proposed, it should be noted that those raters who were led to believe that they would have to justify their ratings in a feedback discussion group exhibited greater convergent validity than those raters who believed that they did not have to justify their ratings (M ICC Justification = .265; M ICC No Justification = .211). Further, raters in the justification condition exhibited lower discriminant validity and lower halo (M ICCs = .247 and .057 respectively) than those raters in the no justification condition (M ICCs = .337 and .096, respectively). However, the halo effects across all conditions were generally low.

To summarize, mixed support was found for the ability of rater training, scale format, and the training x format interaction to influence convergent validity, discriminant validity, and halo across the three rater sources. Specifically, training and the use of the behavioral checklist increased discriminant validity and reduced halo, while rating justification served to reduce discriminant validity. However, contrary to expectations, neither training nor the use of the behavioral checklist enhanced convergent validity.

Table 13 provides only a cursory evaluation of the influence of training, scale format, and rating justification on the quality of ratings provided by the three rater sources. In an attempt to test for the effects of variation in the experimental conditions statistically, procedures set forth by Hedges and Olkin (1983) were employed. These methods can be used to test linear models in research where the dependent variable is a Pearson product moment correlation coefficient. Both Dickinson et al. (1986) and Kavanaugh et al. (1971) believe that ICCs can be treated as having a sampling distribution

approximately the same as the Pearson product moment correlation coefficient. Thus, the Hedges and Olkin (1983) formulas were adopted for use with ICCs in the present study.

Specifically, Hedges and Olkin (1983) use a generalized least squares procedure (see, e.g., Goldberger, 1964) in which the data are analyzed in a regression context with ICCs transformed to Fisher's z scores as the dependent variables. Treatment conditions are the predictor variables (i.e., Training, Format, Justification, Training x Format, Training x Justification, Format x Justification) and their effects are estimated by their beta weights in the regression analysis. This procedure provides a test for the effect of each of the treatments. If the hypothesis of no effects (i.e., all betas equal to zero) is rejected by means of a q statistic, confidence intervals are constructed using Bonferroni inequalities to allow for an examination of the individual treatment effects. In addition, a test for model specification (Q) provides a basis for deciding whether the variation in the transformed ICCs is accounted for by the explanatory variables in the model. Thus, the test for model specification provides a means for evaluating models that explain variation in effect magnitude as a function of experimental conditions (Hedges & Olkin, 1983).

Table 14 provides results of these analyses for convergent validity, discriminant validity, and halo. In each instance the test for model mispecification (Q) did not reject the specification of the analysis of variance model. Separate q tests were calculated for convergent validity, discriminant validity, and halo to test the hypothesis that all betas are equal to zero. As indicated in Table

Table 14

Test for the Effect of Variations in Training, Format, and Rating Justification for Convergent and Discriminant Validity and Halo.

Convergent Validity		
q = 5.09 -- distributed as Chi-square with 7 degrees of freedom.		
Source	Beta	95% Confidence Interval
Grand Mean	0.2430	-0.054 to 0.540
Training (T)	0.0051	-0.292 to 0.302
Format (F)	-0.0290	-0.326 to 0.268
Justification (J)	0.0288	-0.268 to 0.326
T x F	-0.0111	-0.308 to 0.286
T x J	0.0161	-0.281 to 0.313
F x J	0.0065	-0.303 to 0.291
Q = 0.016 -- distributed as Chi-square with 1 degree of freedom.		

Table 14 (Continued)

Discriminant Validity		
q = 29.96* -- distributed as Chi-square with 7 degrees of freedom.		
Source	Beta	95% Confidence Interval
Grand Mean	0.3091*	0.138 to 0.481
Training (T)	0.0685	-0.103 to 0.240
Format (F)	0.1186	-0.053 to 0.290
Justification (J)	-0.0546	-0.226 to 0.117
T x F	0.0565	-0.115 to 0.228
T x J	-0.0284	-0.200 to 0.143
F x J	-0.0250	-0.197 to 0.146
Q = 0.124 -- distributed as Chi-square with 1 degree of freedom.		

Table 14 (Concluded)

Halo		
q = 1.97 -- distributed as Chi-square with 7 degrees of freedom.		
Source	Beta	95% Confidence Interval
Grand Mean	0.0777	-0.133 to 0.287
Training (T)	-0.0021	-0.212 to 0.208
Format (F)	-0.0589	-0.269 to 0.151
Justification (J)	-0.0204	-0.230 to 0.190
T x F	0.0129	-0.197 to 0.223
T x J	0.0311	-0.179 to 0.241
F x J	-0.0309	-0.179 to 0.241
Q = 0.047 -- distributed as Chi-square with 1 degree of freedom.		

Note. The degrees of freedom associated with each ICC value was used to represent N in the development of a source's beta. These degrees of freedom are a conservative estimate of the total number of observations associated with that source.

*p < .05.

14, only the q statistic for discriminant validity was significant. The only confidence interval for discriminant validity that does not contain zero is that for the Grand Mean. This indicates that the ICCs are, as a group, different from zero, but there are no training, format, or rating justification effects. Although the Format beta for discriminant validity approached significance (C.I. = -.053 to .290), its confidence interval contained zero. These results suggest that Hypotheses 1, 2, and 4 were not supported for convergent validity, discriminant validity, or halo. Conclusions regarding these results must be tempered however. It must be noted that the sample sizes were very small (ranging from 9 to 33) greatly reducing the power associated with these tests. Given the lack of power, it is not surprising that significant effects were not found.

Comparison with MTMR Research

Table 15 presents a comparison of the ICC values obtained in the present study to other MTMR studies. In their meta-analysis, Dickinson et al. (1986) identified 28 studies which used rater source as the method. Sixteen of these studies are presented in Table 15. Studies were chosen for inclusion based on their compatability with the present research. That is, the present study was concerned with the quality of ratings exhibited by self-, peer, and observer raters. Past research has shown the greatest discrepancies to exist between self-ratings and other rater sources. Therefore, studies which included self- and/or peer ratings are listed. Table 15 also presents mean ICC values for all 28 rater source studies identified by Dickinson et al. (1986) as a further comparison group. ICC values in the table were computed according to Bartko's (1966) definition (i.e.,

Table 15

Comparison of ICC Values Derived from Previous MTMR Studies.

Study	Convergent Validity	Discriminant Validity	Halo
Tucker, Cline, & Schmitt (1967)			
Study 1	.355	.049	.431
Study 2	.315	.107	.448
Gunderson & Ryman (1971)	.449	.168	.107
Dickinson & Tice (1973)	.179	.072	.273
Orpen (1973)	.322	.121	.044
Borman (1974)	.312	.077	.171
Heneman (1974)	.202	.098	.190
Blackburn & Clark (1975)	.335	.123	.282
Borman, Hough, & Dunnette (1976)	.233	.054	.283
Baird (1977)	.352	.026	.515
Holzbach (1978)			
Study 1	.249	.068	.395
Study 2	.232	.054	.393
Braskamp, Caulley, & Costin (1979)			
Study 1	.217	.146	.176
Study 2	.343	.238	.009
Marsh, Overall, & Kesler (1979)	.179	.294	.167
Marsh (1982)	.129	.301	.151
Mean ICC Value Across Studies (N=16)	.275	.125	.252
Mean ICC Value Across All Studies (N=28)	.289	.104	.256

Table 15 (Concluded)

Present Study (Mean ICC Values)			
Combined	.238	.292	.077
Training	.243	.349	.076
No Training	.233	.235	.078
Graphic Rating Scale	.266	.189	.136
Behavioral Checklist	.210	.394	.018
No Training/Graphic Rating Scale	.251	.178	.150
No Training/Behavioral Checklist	.215	.292	.006
Training/Graphic Rating Scale	.280	.201	.122
Training/Behavioral Checklist	.206	.497	.030

the ratio of a source's variance component to the sum of all relevant variance components).

In general, the convergent validities obtained in the present study are comparable to those obtained in other studies. The average convergent validity as indicated by the ICC value in Table 15 was .238 as compared to .275 for the 16 studies which used self- and/or peer ratings as a rater source. Although the mean ICC value obtained in this study is slightly lower than the 28 study comparison group (.238 compared to .289), both values indicate moderate convergent validity. Further, the mean values in the present study were much higher for discriminant validity and much lower for halo than the two comparison groups. Discriminant validities in other rater source studies tended to be low (M ICCs = .125 and .104), while evidence was found for a moderate amount of discriminant validity in the current study (M ICC = .292). On the other hand, ICC values for halo in the two comparison groups suggest moderate amounts of halo (M ICCs = .252 and .256), while very little halo existed in the present study (M ICC = .077). Overall, the present study indicates comparable convergent validity, less halo, and greater discriminant validity than other rater source studies.

A closer examination of Table 15 reveals where some of the differences occurred in the present study. As noted, most rater source research has used a graphic rating scale and failed to provide rater training. An appropriate point of comparison then is the No Training-Graphic Rating Scale condition used in the present study. A comparison of this group with other studies reveals comparable convergent and discriminant validity (.251 and .178, respectively

compared to .275 and .125 for the 16 study comparison group) and lower halo (.150 compared to .252 and .256). In contrast, rater sources who received training and used the behavioral checklist exhibited much greater discriminant validity (.497 compared to .125 and .104) and less halo (.030 compared to .252 and .256). However, this group also provided less evidence for convergent validity (.206 compared to .275 and .289). The major difference in the present study appears to lie with the use of the behavioral checklist. In every instance that the behavioral checklist was used discriminant validity was higher and halo lower than in other studies. However, although training did not enhance convergent validity as hypothesized, rater sources who received training and used the behavioral checklist achieved moderate levels of convergent validity in addition to exhibiting high discriminant validity and low halo.

Table 15 provided a comparison of the ICC values found in the present study with those of other rater source studies. Unfortunately, no definitive conclusions regarding the influence of training and scale format on the quality of performance ratings made by different rater sources could be made from this descriptive comparison. However, an integration of the present findings with the 28 rater source studies identified by Dickinson et al. (1986) would provide information that can be used to identify the variables that influence convergent validity, discriminant validity, and halo. Further, the problems noted earlier with respect to statistical power can be reduced by synthesizing the results of all rater source research. Consequently, the Hedges and Olkin (1983) procedure described earlier was used to identify the cummulative results of past

rater source studies so that conclusions could be drawn regarding the effects of each of the treatment conditions on convergent validity, discriminant validity, and halo.

Specifically, the 28 rater source studies identified by Dickinson et al. (1986) were combined with the eight experimental conditions in the present study and categorized into four treatment conditions: (1) training/behaviorally-based scales, (2) training/non-behaviorally-based sales, (3) no training/behaviorally-based scales, and (4) no training/non-behaviorally-based scales. The number of studies in each category were 4, 2, 10, and 20 respectively. The behaviorally-based scales included BARS, BES, MSS, and behavioral checklists, while non-behaviorally-based scales included graphic rating scales, summated scales, comparative rating scales, and nomination techniques. The Hedges and Olkin (1983) procedure was used to assess the effects of the treatment conditions (i.e., Training, Type of Format, and Training x Format) on the quality of performance ratings. Table 16 summarizes the results of these analyses for convergent validity, discriminant validity, and halo.

Separate q tests for the hypothesis that the betas are equal to zero were calculated for convergent validity, discriminant validity, and halo. As indicated in Table 16, all three q statistics were significant. With respect to convergent validity, an examination of the beta weights revealed that the confidence intervals for the Grand Mean, Format, and Training x Format effects do not contain zero. The significant beta for the Grand Mean suggests that the ICCs are, as a group, different from zero. Interpretation of the Format effect suggests that studies which used behaviorally-based rating scales

Table 16

Cummulative Test for the Effect of Variations in Training and Format for Rater Source Studies.

Convergent Validity		
q = 224.43** -- distributed as Chi-square with 4 degrees of freedom.		
Source	Beta	95% Confidence Interval
Grand Mean	0.2617*	0.188 to 0.335
Training (T)	-0.0222	-0.086 to 0.042
Format (F)	0.0704*	0.023 to 0.118
T x F	0.0628*	0.005 to 0.120
Q = 26.36 -- distributed as Chi-square with 32 degrees of freedom.		
Discriminant Validity		
q = 253.43** -- distributed as Chi-square with 4 degrees of freedom.		
Source	Beta	95% Confidence Interval
Grand Mean	0.1654*	0.134 to 0.197
Training (T)	0.0507*	0.025 to 0.077
Format (F)	-0.0379*	-0.054 to -0.022
T x F	-0.0187	-0.041 to 0.004
Q = 157.59** -- distributed as Chi-square with 32 degrees of freedom.		

Table 16 (Concluded)

Halo		
q = 245.80** -- distributed as Chi-square with 4 degrees of freedom.		
Source	Beta	95% Confidence Interval
Grand Mean	0.2069*	0.144 to 0.270
Training (T)	-0.0398	-0.096 to 0.017
Format (F)	-0.0523*	-0.095 to -0.010
T x F	0.0080	-0.044 to 0.060
Q = 55.34* -- distributed as Chi-square with 32 degrees of freedom.		

Note. The degrees of freedom associated with each ICC value was used to represent N in the development of a source's beta. These degrees of freedom are a conservative estimate of the total number of observations associated with that source.

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

exhibited greater convergent validity than studies that did not use behaviorally-based scales (M ICCs = .321 and .284, respectively). This finding provides some support for Hypothesis 2 in the present study. That is, it was suggested that rater sources which used the behavioral checklist would exhibit greater convergent validity than rater sources who used the graphic rating scale. While this study's data did not support this hypothesis (see Tables 13 and 14), the results found in Table 16 suggest that when behaviorally-based rating scales are used convergent validity is enhanced. With respect to the significant Training x Format effect, a comparison of the mean ICCs for the four conditions revealed that the no training/behaviorally-based rating scale studies exhibited greater convergent validity (M ICC = .337) than the other three conditions (M ICC Training/Behaviorally-based rating scales = .283; M ICC Training/Non-Behaviorally-based rating scales = .280; M ICC No Training/Non-Behaviorally-based rating scales = .285).

The discriminant validity results in Table 16 indicate significant treatment effects for Training and Type of Format (betas = 0.051 and -0.038, respectively). The mean ICC value for studies that provided rater training was .255, while the mean ICC value in those studies that did not give training was .117. This finding provides support for Hypothesis 1 in the present study. This study failed to find a significant training effect, possibly due to a lack of statistical power caused by the small sample sizes used in the Hedges and Olkin (1983) analysis. However, the discriminant validity results in Table 16 indicate that training does, in fact, influence discriminant validity as hypothesized. Further, the significant beta

weight for the Format effect provides additional support for Hypothesis 2. Discriminant validity was higher in studies where behaviorally-based scales were used (M ICC = .170 compared to .127). Finally, Table 16 indicates a significant Format effect for halo (beta = -0.052) which once again supports Hypothesis 2. Studies which used behaviorally-based scales exhibited less halo than studies that did not use behaviorally-based rating scales (M ICCs = .133 and .247, respectively.

A note of caution is necessary here. The results in Table 16 indicate that the test for model misspecification (Q) for discriminant validity and halo rejected the specification of the analysis of variance model. In the case of misspecified models conclusions made about the effects of the variables in the analyses must be tempered since the estimates of beta may not be consistent. Misspecification of the model is often due to differences in pretreatment controls of unmeasured variables across studies. Given the diversity of conditions under which performance ratings were obtained in the rater source research reviewed, it is not surprising that the model was misspecified. Differences in the rating task, experience of the raters, the different rating scales, and random assignment of subjects to treatments are some of the factors that could have contributed to the model's misspecification.

Leniency Results

Table 17 summarizes the 2 x 2 x 2 x 3 x 4 ANOVA used to test leniency effects in the present study. An examination of Table 17 reveals significant values for the main effects of Rater Source and Dimensions, while significant interactions were found for Training x

Table 17

Analysis of Variance Summary Table Used to Test for Leniency Effects.

Source	df	MS	F-Ratio
Training (T)	1	.03	0.01
Format (F)	1	.01	0.01
Justification (J)	1	5.27	2.28
Rater Source (S)	2	3.25	7.16**
Dimensions (D)	3	12.94	15.98**
Ratees (R)/TxFxJ	88	2.31	No Term
T x F	1	1.12	0.49
T x J	1	.99	0.43
T x S	2	.92	2.04
T x D	3	3.62	4.47**
F x J	1	27.68	11.98**
F x S	2	1.44	3.17*
F x D	3	1.16	1.43
J x S	2	.26	0.56
J x D	3	.95	1.17
S x D	6	.15	0.65
T x F x J	1	.27	0.12
T x F x S	2	1.92	4.23*
T x F x D	3	.35	0.44
T x J x S	2	1.73	3.81*
T x J x D	3	.42	0.52
T x S x D	6	.41	1.84

Table 17 (Concluded)

Source	df	MS	F-Ratio
F x J x S	2	.88	1.94
F x J x D	3	.23	0.29
F x S x D	6	.44	1.95
J x S x D	6	.24	1.09
T x F x J x S	2	5.07	11.19**
T x F x J x D	3	.39	0.48
T x F x S x D	6	.70	3.11*
T x J x S x D	6	.22	0.99
F x J x S x D	6	.25	1.11
T x F x J x S x D	6	.10	0.45
D x R/TxFxJ	264	.81	No Term
S x R/TxFxJ	176	.45	No Term
S x D x R/TxFxJ	528	.22	No Term

Note. Abbreviations are: df = degrees of freedom; MS = Mean squares.

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

Dimension, Format x Justification, Format x Source, Training x Format x Source, Training x Justification x Source, Training x Format x Justification x Source, and Training x Format x Source x Dimension. Hypotheses 1 through 4 proposed that differences in the level of performance ratings would be influenced by training, format, rating justification, and the training x format interaction. A test of these hypotheses requires an examination of the rater source effect and its interaction with these variables.

As indicated in Table 17, there was a main effect for Rater Source ($F(2, 176) = 7.16$, $p < .01$). A Tukey (hsd) post hoc test revealed self-ratings to be more lenient than observer ratings. There was no difference between self- and peer ratings or peer and observer ratings of performance.

Support for Hypothesis 1 (Training) and Hypothesis 3 (Justification) required significant Training x Source and Justification x Source interactions, respectively. These hypotheses were not supported. The Training x Source ($F(2, 176) = 2.04$) and Justification x Source ($F(2, 176) = 0.56$) effects did not influence the level of ratings across rater sources. However, there is support for Hypothesis 2 (Format) as revealed in Table 17. Scale format did influence leniency as indicated by the significant Format x Source interaction ($F(2, 176) = 3.17$, $p < .05$). Tests for simple effects are presented in Table 18. The top half of Table 18 reveals the hypothesized difference among rater sources when the graphic rating scale was used ($F(2, 176) = 9.64$, $p < .01$). A Tukey (hsd) post hoc test indicated that self- and peer rater sources were more lenient than the observer source. As anticipated, no difference among rater

Table 18

Analysis of Variance for Format and Rater Source Simple Effects for the Format x Source Interaction.

Rater Source Simple Effects			
Source	df	MS	F-Ratio
Graphic Rating Scale	2	4.34	9.64**
Behavioral Checklist	2	.25	0.55

Format Simple Effects			
Source	df	MS	F-Ratio
Self	1	1.12	2.49
Peer	1	.16	0.35
Observer	1	1.60	3.55*

Note. The error term for all sources of variation above was the original error term for the Format x Source interaction: S x R/TxFXJ = .45, df = 176. Abbreviations are: df = degrees of freedom; MS = Mean squares.

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

sources was found when the behavioral checklist was used ($F(2, 176) = 0.55$). Further, the bottom half of Table 18 shows that the only difference between rating formats occurred with the observer rater source. In this instance observer raters who used the behavioral checklist were more lenient than observer raters who used the graphic rating scale.

The significant Training x Format x Source interaction in Table 17 provides support for Hypothesis 4. Training and type of format interacted to influence the leniency of ratings across rater sources ($F(2, 176) = 4.23$, $p < .05$). Tests for simple effects were calculated separately for each training condition for the Training x Format x Source interaction and are presented in Table 19. Figure 1 presents a graphic display of the interaction.

An examination of the top half of Table 19 reveals a significant difference among rater sources in the no training condition. Self- and peer ratings were more lenient than observer ratings when no training was received. More specifically, a Tukey (hsd) post hoc test was performed on the significant Format x Source interaction within the no training condition ($F(2, 176) = 7.41$, $p < .01$). As depicted in Figure 1, when no training was provided and self-ratings were made with the graphic rating scale, they were significantly more lenient than ratings provided by any of the three rater sources that used the behavioral checklist without training (i.e., behavioral checklist-self-ratings, behavioral checklist-peer ratings, behavioral checklist-observer ratings). Further, self-ratings made with the graphic rating scale when no training was received were more lenient than those provided by observer raters who did not receive training and used the

Table 19

Analysis of Variance for No Training and Training Simple Effects for the Training x Format x Source Interaction.

No Training Simple Effects			
Source	df	MS	F-Ratio
Format	1	.69	0.30
Source	2	3.67	8.15*
Format x Source	2	3.33	7.41*

Training Simple Effects			
Source	df	MS	F-Ratio
Format	1	.45	0.19
Source	2	.50	1.12
Format x Source	2	.02	0.04

Note. The error term for the Format effect was the original error term for the Training x Format interaction: R/TxFxJ = 2.31, df = 88. The error term for the Source and Format x Source effects was the original error term for the Training x Source and Training x Format x Source interactions: S x R/TxFXJ = .45, df = 176. Abbreviations are: df = degrees of freedom; MS = Mean squares.

*$p < .01$.

graphic rating scale. That is, when raters were trained the rating scale used made little difference in leniency, however, when raters did not receive training the behavioral checklist helped to reduce leniency.

The bottom half of Table 19 presents the simple effects tests for the Training x Format x Source interaction calculated on the training condition. These results indicate that when training was received by all three rater sources no significant differences among rater sources, format, or the format x source interaction occurred. This relationship is also presented pictorially at the bottom of Figure 1. Apparently, the significant Training x Format x Source interaction found in Table 17 was the result of self-raters who used the graphic rating scale and received no training. These ratings were more lenient than those made under most other research conditions. When self-raters were provided with training there was no difference in the level of ratings across rater sources.

Although no specific hypotheses were made with respect to an interaction of rating justification with training or scale format, a significant Training x Justification x Source interaction was found in Table 17 ($F(2, 176) = 3.81$, $p < .05$). Tests for simple effects were calculated separately for each training condition and are presented in Table 20. A graph of these relationships is provided in Figure 2.

As shown in Table 20, results of the simple effects tests for the no training condition revealed that when no training was provided self- and peer ratings were more lenient than observer ratings ($F(2, 176) = 8.15$, $p < .01$). A Tukey (hsd) post hoc test was then performed on the significant Justification x Source interaction found within the

Table 20

Analysis of Variance for No Training and Training Simple Effects for the Training x Justification x Source Interaction.

No Training Simple Effects

Source	df	MS	F-Ratio
Justification	1	5.41	2.34
Source	2	3.67	8.15**
Justification x Source	2	1.41	3.14*

Training Simple Effects

Source	df	MS	F-Ratio
Justification	1	.85	0.37
Source	2	.50	1.12
Justification x Source	2	.57	1.26

Note. The error term for the Justification effect was the original error term for the Training x Justification interaction: R/TxFxJ = 2.31, df = 88. The error term for the Source and Justification x Source effects was the original error term for the Training x Source and Training x Justification x Source interactions: S x R/TxFXJ = .45, df = 176. Abbreviations are: df = degrees of freedom; MS = Mean squares.

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

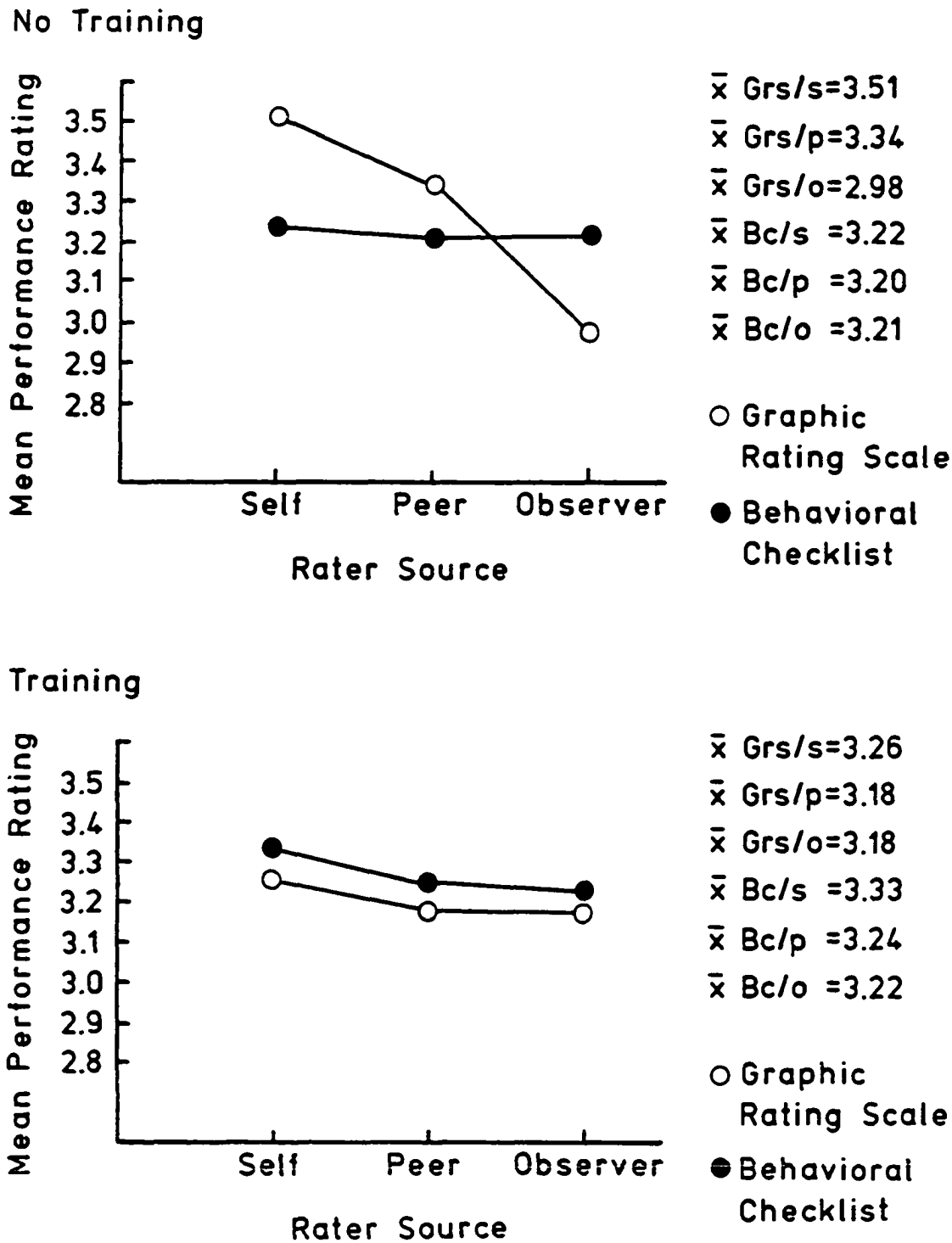

Figure 1

Simple Effects for Training x Format x Source Interaction

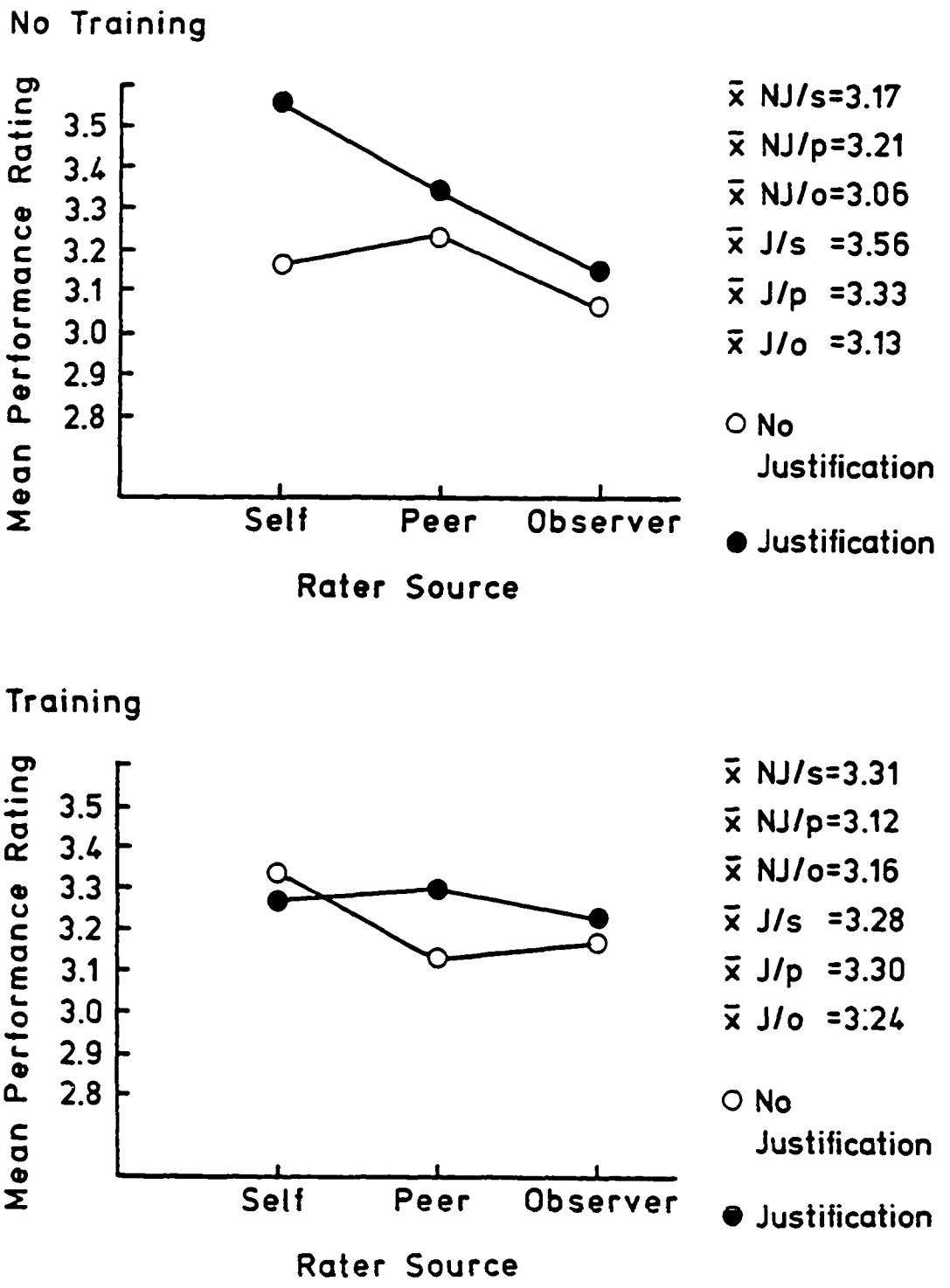

Figure 2

Simple Effects for Training x Justification x Source Interaction

no training condition ($F(2, 176) = 3.14$, $p < .05$). As depicted in Figure 2, self-raters who had to justify their ratings were more lenient than all other conditions. However, when training was provided simple effects tests indicated no differences in the leniency of ratings provided by the three rater sources for the Training x Justification x Source interaction. This result is clearly depicted at the bottom of Figure 2.

A simple effects test was calculated on each training condition for the Training x Format x Justification x Source interaction that was reported to be significant in Table 17. The results of these analyses are presented in Table 21. An interpretation of this four-way interaction is presented only as it provides an additional piece of information that is helpful in understanding the significant Training x Format x Source and Training x Justification x Source interactions.

An examination of Table 21 indicates that when no training was provided a significant Format x Justification x Source interaction was found ($F(2, 176) = 11.18$, $p < .01$). A Tukey (hsd) post hoc test was performed on this signifcant interaction and indicated that self-raters who used the graphic rating scale, received no training, and had to justify their performance ratings, were more lenient than all other possible combinations. Further, peers who used the graphic rating scale, received no training, and had to justify their ratings, were more lenient than: (a) observer raters who used the graphic rating scale without training and who had to justify their ratings, and (b) observer raters who used the graphic rating scale without training and did not have to justify their ratings. In contrast, when

Table 21

Analysis of Variance for No Training and Training Simple Effects for the Training x Format x Justification x Source Interaction.

No Training Simple Effects

Source	df	MS	F-Ratio
Format	1	.69	0.30
Justification	1	5.41	2.34
Source	2	3.67	8.15**
Format x Justification	1	11.22	4.86*
Format x Source	2	3.33	7.41**
Justification x Source	2	1.41	3.14*
Format x Justification x Source	2	5.03	11.18**

Table 21 (Concluded)

Training Simple Effects

Source	df	MS	F-Ratio
Format	1	.45	0.19
Justification	1	.85	0.37
Source	2	.50	1.11
Format x Justification	1	16.74	7.24**
Format x Source	2	.02	0.04
Justification x Source	2	.57	1.26
Format x Justification x Source	2	.92	2.04

Note. The error term for the Format, Justification, and Format x Justification effects was the original error term for the Training x Format, Training x Justification, and Training x Format x Justification interactions: R/TxFxJ = 2.31, df = 88. The error term for the Source, Format x Source, Justification x Source, and Format x Justification x Source effects was the original error term for the Training x Source, Training x Format x Source, Training x Justification x Source, and Training x Format x Justification x Source interactions: S x R/TxFXJ = .45, df = 176. Abbreviations are: df = degrees of freedom; MS = Mean squares.

*$p < .05$. **$p < .01$.

training was provided to rater sources, no difference in the level of ratings was found across rater sources or between justification conditions regardless of which format was used.

Table 17 also revealed significant effects for Dimensions and the Training x Dimension, Format x Justification, and Training x Format x Source x Dimension interactions. A Tukey (hsd) post hoc test on the Dimension effect revealed that the dimensions of problem analysis, problem solution, and sensitivity were rated more leniently than the dimension of persuasiveness.

Table 22 provides the simple effects tests for the Training x Dimension interaction, while Table 23 presents similar analyses for the Format x Justification interaction. The analyses in Table 23 show a significant difference among dimensions when no training was provided ($F(3, 264) = 17.86$, $p < .01$). This difference, however, did not occur when raters were provided with training ($F(3, 264) = 2.58$). A Tukey (hsd) post hoc test revealed that when no training was provided ratings on problem analysis, problem solution, and sensitivity were more lenient than ratings on persuasiveness. Further, results of the training simple effects tests revealed a significant difference between training conditions for both persuasiveness and sensitivity. With respect to persuasiveness, raters in the training condition were more lenient than raters in the no training condition. When rating sensitivity, raters in the no training condition were more lenient than raters who received training.

Simple effects tests for the Format x Justification condition (Table 23) indicated a difference between justification conditions for raters who used the graphic rating scale ($F(1, 88) = 12.36$, $p < .01$).

Table 22

Analysis of Variance for Training and Dimension Simple Effects for the Training x Dimension Interaction.

Dimension Simple Effects			
Source	df	MS	F-Ratio
No Training	3	14.47	17.86**
Training	3	2.09	2.58

Training Simple Effects			
Source	df	MS	F-Ratio
Problem Analysis	1	.06	0.07
Problem Solution	1	.01	0.01
Persuasiveness	1	4.28	5.28*
Sensitivity	1	6.54	8.07**

Note. The error term for all sources of variation above was the original error term for the Training x Dimension interaction: D x R/TxFXJ = .81, df = 264. Abbreviations are: df = degrees of freedom; MS = Mean squares.

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

Table 23

Analysis of Variance for Format and Justification Simple Effects for the Format x Justification Interaction.

Format Simple Effects			
Source	df	MS	F-Ratio
Graphic Rating Scale	1	28.55	12.36**
Behavioral Checklist	1	4.40	1.90
Justification Simple Effects			
Source	df	MS	F-Ratio
No Justification	1	13.26	5.74*
Justification	1	14.44	6.25*

Note. The error term for all sources of variation above was the original error term for the Format x Justification interaction: R/TxFXJ = 2.31, df = 88. Abbreviations are: df = degrees of freedom; MS = Mean squares.

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

Raters who used the graphic rating scale and expected to have to justify their performance ratings were more lenient than raters who used the graphic rating scale but did not have to justify their ratings. There was also a significant difference between formats for both the justification and no justification conditions. In the no justification condition, raters who used the behavioral checklist were more lenient than those who used the graphic rating scale. On the other hand, in the justification condition, raters who used the graphic rating scale were more lenient than those who used the behavioral checklist.

Finally, simple effects tests were calculated on each training condition for the Training x Format x Source x Dimension interaction that was found in Table 17. The results of these analyses are presented in Table 24. An examination of this table indicates a significant Format x Source x Dimension interaction for both the training and no training conditions ($F(6, 528) = 2.60$, $p < .05$ and $F(6, 528) = 2.54$, $p < .05$, respectively). To examine these three-way interactions more closely simple effects tests were calculated for each dimension for the Format x Source x Dimension interaction. These results are presented in Table 25.

Table 25 indicates that rater sources differed in their ratings of the performance dimensions under various treatment conditions. Tukey (hsd) post hoc tests revealed that when no training was provided self-ratings made with the graphic rating scale were significantly higher on the dimension of problem analysis than observers who used the graphic rating scale without training, and self- and peer raters who used the behavioral checklist with no training. Differences were

Table 24

Analysis of Variance for No Training and Training Simple Effects for the Training x Format x Source x Dimension Interaction.

No Training Simple Effects

Source	df	MS	F-Ratio
Format	1	.69	0.30
Source	2	3.67	8.15**
Dimension	3	14.47	17.86**
Format x Source	2	3.33	7.41**
Format x Dimension	3	.15	0.19
Source x Dimension	6	.36	1.63
Format x Source x Dimension	6	.57	2.60*

Table 24 (Concluded)

Training Simple Effects

Source	df	MS	F-Ratio
Format	1	.45	0.19
Source	2	.50	1.12
Dimension	3	2.09	2.58
Format x Source	2	.02	0.04
Format x Dimension	3	1.36	1.68
Source x Dimension	6	.20	0.89
Format x Source x Dimension	6	.56	2.54*

Note. The error term for the Format effect was the original error term for the Training x Format interaction: R/TxFxJ = 2.31, df = 88. The error term for the Source and Format x Source effects was the original error term for the Training x Source and Training x Format x Source interactions: S x R/TxFxJ = .45, df = 176. The error term for the Dimension and Format x Dimension effects was the original error term for the Training x Dimension and Training x Format x Dimension interactions: D x R/TxFxJ = .81, df = 264. The error term for the Source x Dimension and Format x Source x Dimension effects was the original error term for the Training x Source x Dimension and Training x Format x Source x Dimension interactions: S x D x R/TxFxJ = .22, df = 528. Abbreviations are: df = degrees of freedom; MS = Mean squares. *p < .05. **p < .01.

Table 25

Analysis of Variance for Dimensions within Training Conditions for the Format x Source x Dimension Interaction.

No Training Simple Effects

Problem Analysis

Source	df	MS	F-Ratio
Format	1	.64	0.79
Source	2	1.67	7.62**
Format x Source	2	2.16	9.82**

Problem Solution

Source	df	MS	F-Ratio
Format	1	.49	0.60
Source	2	1.71	7.76**
Format x Source	2	2.19	9.94**

Persuasiveness

Source	df	MS	F-Ratio
Format	1	.01	0.01
Source	2	1.05	4.77**
Format x Source	2	.17	0.81

Table 25 (Continued)

Sensitivity			
Source	df	MS	F-Ratio
Format	1	.00	0.00
Source	2	.31	1.42
Format x Source	2	.52	2.38

Training Simple Effects			
Problem Analysis			
Source	df	MS	F-Ratio
Format	1	.61	0.76
Source	2	.63	2.86
Format x Source	2	.05	0.22
Problem Solution			
Source	df	MS	F-Ratio
Format	1	.11	0.13
Source	2	.03	0.12
Format x Source	2	.12	0.55

Table 25 (Concluded)

Persuasiveness			
Source	df	MS	F-Ratio
Format	1	3.46	4.27*
Source	2	.01	0.06
Format x Source	2	1.05	4.78**
Sensitivity			
Source	df	MS	F-Ratio
Format	1	.35	0.43
Source	2	.43	1.94
Format x Source	2	.47	2.16

Note. The error term for the Format effect was the original error term for the Training x Format x Dimension interaction: D x R/TxFxJ = .81, df = 264. The error term for the Source and Format x Source effects was the original error term for the Training x Source x Dimension and Training x Format x Source x Dimension interactions: S x D x R/TxFxJ = .22, df = 528. Abbreviations are: df = degrees of freedom; MS = Mean squares.

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

also found for problem solution and persuasiveness when no training was provided. For problem solution, self-raters who used the graphic rating scale were more lenient than observers who used the graphic rating scale and all rater sources that used the behavioral checklist, while peers who used the graphic rating scale were more lenient than observers who used the graphic rating scale. For persuasiveness, self- and peer raters were significantly more lenient than observer raters. On the other hand, when training was provided, the only rater source difference that occurred was for the dimension of persuasiveness. Self- and observer raters who used the behavioral checklist were more lenient than observer raters who used the graphic rating scale.

Summary of Results

The principle objective of the present study was to examine the influence of rater training, scale format, and rating justification on the quality of performance ratings (i.e., convergent validity, discriminant validity, halo, leniency) exhibited by three rater sources. The results obtained in this study, while similar in some respects to those reported elsewhere, contain several important differences. In general, the degree of agreement among the rater sources (convergent validity) found in the present study was comparable to that reported in other studies. Further, the degree of discriminant validity was larger and the halo effect smaller than that reported elsewhere. When one examines the experimental conditions separately, however, comparisons with other research becomes more distinct. As noted in Table 15, the quality of performance ratings exhibited by rater sources who did not receive training and used the

graphic rating scale was similar to that of research reported elsewhere. In contrast, rater sources that received training and used the behavioral checklist exhibited lower convergent validity but much greater discriminant validity and less halo than that reported in other studies. With respect to leniency, the level of ratings across different rater sources was affected by the variables of interest. Specifically, training and the use of the behavioral checklist helped to reduce leniency in self-ratings in those situations where raters had to justify their performance ratings.

IV. DISCUSSION

Prior to this study very little was known about the combined influence of rater training, scale format, and rating justification on the quality of performance ratings provided by different rater sources. Four hypotheses were proposed concerning the influence of these variables on self-, peer, and observer raters. Taken as a whole, the data indicate that the rater sources were differentially influenced by these variables. The discussion section focuses on each hypothesis, provides explanations for the results, and integrates the findings of this study with previous research in this area.

Convergent Validity, Discriminant Validity, and Halo

Rater Training. A number of studies have demonstrated that rater training programs can improve the effectiveness of at least some aspects of the performance rating process (e.g., Borman, 1979; Dickinson & Silverhart, 1986; Fay & Latham, 1982; McIntyre et al., 1984; Pulakos, 1984). The present study hypothesized that the absence of a shared frame-of-reference would tend to exaggerate rater discrepancies since each rater must then supply his of her own frame-of-reference. It was believed that rater training would help different rater sources develop a common frame-of-reference for evaluating performance which would, in turn, improve the quality of performance ratings (i.e., convergent and discriminant validity, halo) across rater sources. Mixed support was found for this hypothesis. Rater sources who received training exhibited greater degrees of

discriminant validity than rater sources in the no training condition (see Table 13). In addition, discriminant validity in the training group was greater than that reported elsewhere (see Table 15). Further, when the data in this study were combined with other rater source studies, a significant training effect was revealed for discriminant validity (see Table 16). Contrary to Hypothesis 1, however, there was no difference between training conditions for convergent validity (M ICC Training = .243; M ICC No Training = .238) and halo (M ICC Training = .076; M ICC No Training = .077).

The success of the training program in improving discriminant validity (M Training ICC = .349; M No Training ICC = .235) can be traced to the detailed practice and feedback provided to raters which helped different rater sources to develop common standards of effective performance in the role play exercise. This portion of the training, Performance Dimension Training (Smith, 1986), familiarized rater sources with the dimensions by which performance was rated, thus, improving discriminant validity. Dickinson et al. (1986) recommended that one way to improve discriminant validity was to provide rater training. Their meta-analysis, however, was not able to make recommendations on specific training program content. The findings reported here are similar to those reported in other studies which have used Performance Dimension Training to improve the quality of performance ratings (e.g., Fay & Latham, 1982; Pulakos, 1984; Pursell et al., 1980). The fact that discriminant validity in the present study was enhanced with the training of inexperienced raters is encouraging. The accumulation of these findings suggest that researchers attempting to improve discriminant validity may wish to

incorporate Performance Dimension Training into their training programs.

Several possibilities exist to help explain why rater training failed to enhance convergent validity as hypothesized. First, the raters used in the present study were all college undergraduate students (M Age = 24). For most, this was their first exposure to performance ratings. Expectations for high convergent validity among raters who have never provided performance ratings may have been unrealistic. To the extent that experienced raters had been used, the training provided might have had a greater impact on convergent validity.

In addition, the Dickinson et al. (1986) meta-analysis identified three factors negatively correlated with convergent validity that were present in this study. The number of items per dimension correlated negatively with convergent validity (r = -.32). Further, these authors found that performance ratings made in an academic environment versus an organizational environment had a correlation with convergent validity of -.37, while the use of students as raters had a correlation of -.42. It was hoped that providing rater training to raters would overcome these limitations. A comparison of the convergent validities found in the training group to those reported elsewhere (M ICC = .243 compared to .275 for the 16 study comparison group) would suggest that training did not influence convergent validity as anticipated. However, given the three factors present in this study that Dickinson et al. (1986) found to be negatively correlated with convergent validity, the finding of comparable convergent validity is actually support for the effectiveness of the

training program to improve convergent validity (Hypothesis 1). That is, the training program was able to overcome the negative factors present in this study (i.e., student raters, academic setting, large number of items per dimension) to achieve moderate convergent validity. Future research should attempt to replicate this study in an organizational setting with more typical raters, to determine if rater training can enhance convergent validity in different rater sources beyond that found here.

An examination of the training program content provides some further insights into why convergent validity was not significantly higher for those in the training condition. As noted, the training program combined Performance Dimension Training and Performance Standards Training (Smith, 1986). Performance Dimension Training is designed to familiarize raters with the performance dimensions to be rated. This was accomplished through extensive practice and feedback on both the dimensions and the rating scales. The effectiveness of this training component is evident in the high ICC value for discriminant validity and the low degree of halo found for raters in the training condition. However, the effectiveness of the Performance Standards Training component is questionable.

The goal of Performance Standards Training is to aid raters in developing standards for effective performance that are congruent with expert raters (Smith, 1986). This is achieved by presenting samples of job performance to trainees along with the appropriate or "true" score assigned to the performance dimension by trained experts. The training program used in the present study provided raters with an opportunity to discuss what particular ratee behaviors led them to

their practice ratings, but no "true" scores were available for which raters could compare their practice ratings. That is, the training program used here was negligent in providing raters with an important component of Performance Standards Training, behavioral rationales for ratings given by expert raters coupled with the appropriate "true" score. Raters in this study could only compare their practice ratings with ratings provided by the experimenter and the ratings of others in their training group. This could explain why convergent validity was not improved in the training group as hypothesized. That is, if true scores had been presented, higher convergent validity might have resulted. Future research should examine this possibility.

An alternative explanation for the lack of enhanced convergent validity in the training condition is that each rater source may have actually been tapping a unique aspect of the ratee's performance. Borman (1974) argued that raters from different organizational levels have different orientations to the job being rated and are likely to observe different job behaviors. As several researchers suggest, it is possible that ratings made from different rater sources are equally valid despite relatively low degrees of agreement (Dunnette & Borman, 1979; Landy & Farr, 1980). The convergent validity results reported here and in other rater source research suggests that if organizations and researchers are interested in obtaining accurate assessments of performance, they should employ multiple rater sources in the appraisal process. Further, rater sources should only assess dimensions of performance that directly affect them (Kavanaugh et al., 1986). If incumbents, peers, subordinates, supervisors, and others observe work performance under different circumstances or even

perceive the same performance differently, their separate perceptions of the ratees' performance provide unique information. Multiple rater sources may be needed to increase the likelihood that all aspects of work performance are included in the appraisal process since the judgments of a single rater source are limited.

The research direction seems clear. If the goal of performance measurement is to assess job performance with minimal criterion deficiency and maximum accuracy, then research needs to be conducted to identify which rater sources provide high-quality ratings on which performance dimensions. It has been suggested that supervisors may provide better ratings on technical dimensions, and peers may provide useful information on interpersonal dimensions (Dickinson et al., 1986). Others have hypothesized that self-ratings can provide good measures of ability (Kavanaugh et al., 1986), while subordinates may be in the best position to evaluate performance on such dimensions as delegation and work direction since they are able to directly observe managers' performance in these areas (Mount, 1984). The present study found that self-, peer, and observer rater sources differed in their ratings of performance dimensions (see Tables 24 and 25). For example, when rating problem solution, self raters who received no training and used the graphic rating scale, were more lenient than observer raters who used the graphic rating scale and all three sources that used the behavioral checklist.

It will be necessary to determine in subsequent research, for each rater source, which part of the criterion space it can best measure if a multiple-method approach to the assessment of job performance is desired. It will also be necessary to determine under

what conditions high-quality ratings occur for different rater sources. To address this issue, the MTMR design could be extended to include multiple sources, different training conditions, multiple formats, and different types of performance dimensionss (e.g., technical, interpersonal, abilities).

Each of the explanations just provided is a plausible argument for why rater training did not improve convergent validity. It is important to note, however, that the inability of training to improve convergent validity does not mean that the quality of performance ratings in the training condition was poor. The present study found moderate convergent validity for rater sources who received training. In fact, the degree of convergent validity found in the training group was comparable to that found elsewhere. In addition, at the same time that training was "maintaining" convergent validity, discriminant validity was improved and halo reduced. Dickinson et al. (1986) found the intercorrelations among ICC values for convergent validity, discriminant validity, and halo to be negatively correlated (convergent validity and discriminant validity r = -.16, convergent validity and method bias r = -.35, discriminant validity and method bias r = -.56). Consequently, one would not expect to improve all three variables at the same time. Therefore, the fact that rater training was able to maintain a moderate level of convergent validity while improving discriminant validity and reducing halo suggests that the overall quality of ratings was enhanced with rater training.

Rating Format. A review of the rater source research revealed that most studies examining the psychometric properties of different rater sources had used some type of graphic rating scale (e.g.,

Heneman, 1974; Holzbach, 1978; Klimoski & London, 1974; Schneier & Beatty, 1978; Tsui, 1983; Tsui & Ohlott, 1986). It was suggested that the predominant use of the graphic rating scale may have contributed to the poor agreement, low discriminant validity, and high rater bias typically found in the research.

Tests of significance for format ICCs reported in Table 14 did not reveal any significant betas, although the format beta for discriminant validity approached significance. Conclusions based on these analyses suggest that Hypothesis 2 was not confirmed. However, as noted in the results section, the statistical power associated with these tests was low due to the small sample sizes in each experimental condition. By synthesizing the results of all rater source research this problem was alleviated. The results of these analyses (see Table 16) revealed significant Format effects for convergent validity, discriminant validity, and halo. Unfortunately, these analyses grouped several different types of behavioral scales (e.g., BARS, BES, MSS, checklists) into one category, thus preventing a direct comparison of the behavioral checklist with the graphic rating scale. These analyses do suggest, however, that behaviorally-based rating scales produce higher-quality performance ratings. Therefore, it is possible to generalize to the results in the present study when interpreting Format effects.

Tables 13 and 15 indicate that discriminant validity was high in those situations where the behavioral checklist was used (M ICC = .394) and low when the graphic rating scale was used (M ICC = .189). Further, ICC values for halo where higher among graphic rating scale users when compared to raters who used the behavioral checklist (M ICC

= .136 compared to .018). These findings are similar to those reported in Table 16 for the combined samples. Discriminant validity may have been higher and the halo effect lower for rater sources who used the behavioral checklist because the items on the checklist were selected in such a way that maximized their uniqueness, in contrast to the global impressions that were required of rater sources who used the graphic rating scale.

A close look at the characteristics of the behavioral checklist suggests further explanations for why the differences found between the formats for discriminant validity and halo are not surprising. First, the behavioral checklist used in the present study was the product of an extensive systematic developmental process (see Campbell, 1986). This process helped to insure that the performance dimensions were conceptually independent. Non-independent dimensions would have resulted in high intercorrelations between dimensions and a low degree of discriminant validity similar to that found with the graphic rating scale. In addition, the involvement of experts in the development of rating scales, as was the case with the checklist, has been shown to reduce method bias (halo) (Dickinson et al., 1986).

An additional property of the behavioral checklist that may have led to greater discriminant validity and lower halo than that found with the graphic rating scale was the ability to obtain multiple ratings for each performance dimension on the checklist as opposed to the single rating per dimension obtained with the graphic rating scale. The multiple ratings made for each dimension were averaged for each ratee to obtain a measure for that dimension. A number of studies have indicated that the average of ratings is more reliable

than a single rating (French & Bell, 1978; Latham & Wexley, 1981). This averaging process may have resulted in more reliable dimension ratings than those obtained with the graphic rating scale. This possibility is further supported by the findings of Dickinson et al. (1986). These authors reported that the greater the number of ratings per dimension, the lower the method bias and the greater the discriminant validity. Apparently, the additional ratings per dimension helped raters to focus on ratee differences and increased their ability to discriminate among ratees.

Explanations for the inability of the behavioral checklist to enhance convergent validity as hypothesized may be related to the content of the items on the checklist and the method of item selection. The items used on the behavioral checklist had very little redundancy or overlap since they were chosen on the basis of rigorous statistical analyses. It is possible that the existence of only moderate convergent validity is attributable to the specificity of the item content which decreased the likelihood that raters would observe all relevant behaviors over the course of the entire role play exercise. While specificity is a desirable attribute in a checklist, it quickly becomes unmanageable in those situations where the anticipated behaviors are not constrained by the nature of the performance task. As the range of possible behaviors increases, specificity requires an increasing number of items. Discrepancies across individual raters in what behaviors are processed would reduce convergent validity.

Related to this issue is the actual number of items that were on the checklist. Each of the four performance dimensions had 15 items.

Dickinson et al. (1986) reported that an outcome associated with a greater number of ratings per dimension was lower convergent validity (r = -.32). It is possible that the large number of items used on the checklist in the present study adversely affected convergent validity. An attempt was made to minimize this potential problem in the training condition by encouraging raters to take notes during the role play videotape. Unfortunately, the quick pace of the role play exercise, combined with the inability of raters to view the videotape more than once, placed limitations on the effectiveness of this procedure.

Training x Format Interaction. The findings reported here and in the Dickinson et al. (1986) meta-analysis present a stumbling block for researchers and practitioners. The number of items on a rating scale appears to involve a tradeoff between convergent and discriminant validity. While a larger number of items tends to be related to higher discriminant validity (r = .63), it is also related to lower convergent validity (r = -.32). This presents a dilemma to the researcher who is trying to develop a construct valid rating scale. Traditionally, investigators interested in rater source research have found it more difficult to establish high discriminant validity and low method bias. At what point do different degrees of convergent validity, discriminant validity, and halo become acceptable? Should researchers accept lower levels of convergent validity if high discriminant validity and low halo can be attained?

The proposed integration of rater training and scale format is a step toward resolving this issue. Specifically, this study hypothesized that when rater sources used the behavioral checklist and received training, the result would be high convergent and

discriminant validity and low halo. This belief was predicated on the success of recent rater training programs and the inherent characteristics of the behavioral checklist. The results indicated moderate support for this hypothesis. Rater sources who received training and used the behavioral checklist had higher discriminant validity (M ICC = .497) than rater sources who provided performance ratings with the graphic rating scale without training (M ICC = .178). Further, ICC values for halo were generally lower in those instances when training was provided and the behavioral checklist used (M ICC = .030) than in those situations when the graphic rating scale was used without training (M ICC = .150). These findings support the Training x Format interaction hypothesized.

Contrary to this hypothesis, the lowest ICC values for convergent validity were found among those raters who received training and used the behavioral checklist. This does not mean, however, that the training x format condition did not produce high-quality ratings. In fact, the opposite can be argued.

Discriminant validity reflects the differential ordering of the ratees due to the amounts of the traits demonstrated by the ratees. This outcome is always desirable as work performance is multidimensional and ratees should be expected to differ in their rank-ordering from dimension to dimension (Dickinson et al., 1986). Halo, on the other hand, reflects the differential ordering of the ratees by the sources used to obtain the ratings. This bias is undesirable because the differential ordering of ratees should be due to individual differences in the amounts of the traits demonstrated by the ratees and not due to the sources used to make the ratings

(Dickinson et al., 1986). As noted, rater source research has found it difficult to demonstrate high discriminant validity and low halo. Consequently, the fact that the Training-Behavioral Checklist condition maintained a moderate level of convergent validity while improving discriminant validity and reducing halo suggests that this combination of conditions may be important to researchers trying to develop construct valid rating systems. The training component apparently allowed rater sources to establish a common frame-of-reference that helped overcome the characteristics of the checklist that contribute to low convergent validity (i.e., a large number of ratings per dimension). The result was an improved rating system with moderate convergent validity, high discriminant validity, and low halo.

Based on these results it is believed that research integrating rater training and scale format deserves further attention especially given the problems encountered in the present study with Performance Standards Training. The important research question that must be addressed is whether or not training should focus on observation skills, performance dimensions, performance standards, the rating scale, or some combination of these. This study provided initial insights in this regard. The results reported here clearly document the ability of the behavioral checklist to improve discriminant validity and reduce halo. The use of a checklist with the appropriate rater training program may help improve convergent validity beyond that found in the present study while maintaining a high level of discriminant validity and a low degree of halo.

Rating Justification. No specific hypotheses were proposed with

respect to the influence that rating justification would have on the construct validity (i.e., convergent and discriminant validity, halo) of ratings provided by different rater sources. A number of studies have investigated the impact of the intended use of performance ratings on psychometric properties (e.g., McIntyre et al., 1984; Sharon & Bartlett, 1969; Zedeck & Cascio, 1982). These studies have found that ratings are more lenient under conditions of administrative use than under conditions of research use. However, prior to this study no research had examined the effects of rating justification on convergent validity, discriminant validity, and halo. The results reported here indicated that rater sources who were led to believe that they would have to justify their ratings exhibited slightly higher degrees of convergent validity than those rater sources who believed that they did not have to justify their ratings (M ICCs = .265 compared to .211). In addition, raters in the justification condition exhibited lower levels of discriminant validity and halo (M ICC = .247 and .057, respectively) than those rater sources in the no justification condition (M ICC = .337 and .096, respectively).

While no appreciable differences were found for convergent validity and halo, the difference between the justification conditions for discriminant validity suggests that the quality of performance ratings may be affected when raters are aware that they will have to provide the ratee with face-to-face feedback. This finding has important practical implications. That is, the ability of an organization to differentiate among employees for promotion, training, salary increases, etc. is hindered when discriminant validity is low. This is especially true when the purpose of the

performance rating is developmental.

It is possible that raters were reluctant to provide low ratings on some dimensions, because they felt incapable of giving negative feedback. This would reduce discriminant validity across dimensions. A potential solution to this problem may be to provide feedback training to raters if the purpose of the rating is direct feedback. A feedback training component was included in the training program in the present study to aid raters in preparing for the face-to-face feedback discussion group. It was believed that if raters were aware of certain basic characteristics of effective feedback discussions (e.g., the need to observe performance carefully, the need to be specific, the need to focus on behaviors) they would be more confident entering the feedback discussion, and hence, provide higher quality ratings. Evidence supporting this hypothesis was presented in Table 13. Raters in the justification condition who received training exhibited a greater degree of discriminant validity than rater sources in the justification condition who were not provided with training (M ICC = .283 compared to .211). This finding is clearly evident when the Training-Behavioral Checklist-Justification condition is compared to the No Training-Behavioral Checklist-Justification condition (ICC = .400 compared to .266). Unfortunately, the present study is incapable of determining if this difference was the result of the feedback training component or rater training in general. Future research must manipulate the feedback component of training to answer this question. A simple 2 x 2 design could be used with two feedback training conditions (feedback training, no feedback training) and two levels of justification (justify, not justify). In addition, it is recommended

that a more intense feedback component be provided. This might include an in-depth lecture as well as a role play exercise which provides raters with an opportunity to practice their feedback skills.

Leniency

Previous research has found self-ratings to be more lenient than supervisor and peer ratings (e.g., Holzbach, 1978; Klimoski & London, 1974; Mascitti, 1978; Thornton, 1968). The present study proposed four hypotheses concerning leniency in the performance ratings provided by the three rater sources. Specifically, it was suggested that training, scale format, rating justification, and the training x format interaction would influence the level of ratings across self-, peer, and observer rating sources. A test of these hypotheses required an examination of the rater source effect and its interaction with these variables. Results of these analyses were presented in Tables 17 through 25.

Rating Format and Training x Format Interaction. With respect to leniency, the hypotheses that rating format and the training x format interaction would influence the level of performance ratings across the three rater sources were confirmed. Ratings made with the graphic rating scale were more lenient than those made with the behavioral checklist (Hypothesis 2). A Tukey (hsd) post hoc test revealed that when the graphic rating scale was used, self- and peer ratings were higher than observer ratings. However, when the behavioral checklist was used no difference in the level of ratings across the three sources occurred.

A significant training x format x source interaction provided support for Hypothesis 4. Tests for simple effects calculated on each

training condition (Table 20) revealed that when no training was provided and self-ratings were made with the graphic rating scale, they were more lenient than ratings provided by peers and observers who used the graphic rating scale without training. In addition, these ratings were more lenient than self-, peer, or observer ratings made with the behavioral checklist without the aid of training. However, when training was given, no significant effects for leniency occurred for the rater sources regardless of which format was used. These simple effect tests indicate that the training x format x source interaction was the result of lenient self-ratings made with the graphic rating scale when no training was provided.

The finding that self-ratings were more lenient when the graphic rating scale was used and no training was provided is strikingly similar to the results of previous rater source studies (e.g., Holzbach, 1978; Mount, 1984; Thornton, 1968; Tsui, 1983). Researchers have cautioned practitioners to use self-ratings carefully, because it is believed that individuals have a significantly different view of their own performance than that held by other sources (e.g., Borman, 1974; Thornton, 1980). This study, however, has shown that leniency, defined as a significant difference in the level of ratings across sources, is affected by such variables as rater training and scale format. If untrained supervisors commit rating errors such as leniency and halo, rater training is recommended. To expect individuals to evaluate their own performance accurately without training is unrealistic. The results of this study suggest that by providing training, leniency in self-ratings can be reduced.

It is important to remember that training, in and of itself, did

not reduce leniency across the three rater sources. When ratings were made with the graphic rating scale, they were more lenient than ratings made with the behavioral checklist. Differences between rating formats similar to this have been reported elsewhere. In comparing behaviorally anchored rating scales (BARS) to numerically anchored rating scales for leniency, Mascitti (1978) found self- and peer ratings obtained on the numerical scale were more lenient than ratings obtained on the BARS. Saal and Landy (1977), on the other hand, found peer ratings for police officers with a mixed standard scale to result in less leniency than ratings on a BARS for both supervisors and peers.

Two characteristics associated with the behavioral checklist help to explain why ratings on the graphic rating scale were more lenient. First, raters completing the checklist for a given ratee were not required to "evaluate" the individual's performance but were simply asked to check those behaviors on the checklist that were observed. In contrast, the graphic rating scale required raters to view an episode of performance and, based on their observations, evaluate the performance of the ratee on a specified scale from less than acceptable to more than acceptable. That is, the behavioral checklist required the rater to function less as a judge and more as an observer of behavior than the graphic rating scale.

Secondly, the graphic rating scale presented raters with descriptions of different levels of "goodness of performance" for each dimension, and then asked raters to select the level of performance that best described the ratee on that dimension. That is, graphic rating scale users were presented with an "order of merit continuum"

and were fully aware of the rating they were giving to a ratee as they circled the number which they believed accurately represented the performance of the individual on that dimension. Raters who used the behavioral checklist, on the other hand, were unaware of the scale values of each behavioral item they were checking. Although each item on the checklist was assigned a scale value from 1 to 5, this value was unknown to the rater who simply checked a behavior if it occurred. Scoring was completed by the experimenter after all ratings had been gathered. Therefore, raters were prevented from "knowing" what level of rating they gave to a particular ratee. This characteristic of the behavioral checklist reduced the possibility that raters would be able to form a clear picture of an "order of merit" continuum for a dimension rating.

Therefore, the behavioral checklist appears to be a logical approach to the reduction of leniency. By asking the rater to simply check a behavior if it is observed, as opposed to asking the rater to "evaluate" the performance of the ratee on the dimension, and by disguising the scale value of each item, the behavioral checklist would appear to present an obstacle to the rater who, knowingly or unknowingly, rates all individuals "high."

Rater Training and Rating Justification. The hypotheses that the main effects of training and rating justification would influence leniency were not supported in the present research. However, a significant training x format x justification x source interaction was found (see Table 17). Tests for simple effects were presented in Table 21. These results indicated that self-raters who used the graphic rating scale, received no training, and had to justify their

performance ratings, were more lenient than all other possible combinations. In addition, peers who used the graphic rating scale, received no training, and had to justify their ratings, were more lenient than two conditions: observer raters who used the graphic rating scale without training and who had to justify their ratings, and observer raters who used the graphic rating scale without training and who did not have to justify their ratings. In contrast, when training was provided to rater sources, no difference in the level of ratings was found across rater sources or between justification conditions regardless of which format was used.

Research by Stockford and Bissel (1949) supports these findings. These authors found that supervisors who had to explain their performance ratings to their subordinates rated them more leniently than when they did not have to explain them. In addition, no training was provided in the Stockford and Bissel (1949) research, thus, the conditions in the Stockford and Bissel (1949) study closely approximate the No Training-Justification conditions in this study. As noted in the introduction, the implications that "justification" may have on performance appraisal ratings for an organization are considerable. Inflated performance ratings, caused by the influence of justification, are inaccurate and hinders an organization's ability to differentiate among employees for promotions, training, and salary increases. It may be that raters inflated their ratings because they did not want the experience of giving negative feedback to individuals. This could account for the lenient ratings in the justification condition.

Therefore, the finding that no differences in the level of

ratings occurred across rater sources or between justification conditions when training was provided has important practical implications. This suggests that leniency errors in self-ratings can be controlled with training, and is consistent with previous rater training studies that have been successful in reducing leniency in other rater sources (e.g., Bernardin & Pence, 1980; Fay & Latham, 1982; Pulakos, 1984). Self-ratings are faced with an uncertain future as a bona fide method of performance assessment. Inconsistent findings with supervisor-self agreement and inflated ratings have increasingly led to expressions of reservation regarding the practical utility of self-ratings (cf. Thornton, 1980). Whereas the value of self-ratings as vehicles for personal development is typically emphasized, the potential contribution to administrative requirements (e.g., compensation administration, test validation) has been seriously questioned (cf. Cummings & Schwab, 1973). The present study, in and of itself, does not signal a drastic reversal of this trend. However, it does suggest that leniency in self-ratings may be controlled in a fashion similar to that which has been successful with other rater sources (e.g., supervisors, assessment center raters).

Limitations

The conclusions and generalizations of any research study are limited by certain methodological and statistical constraints. Here, generalizations about the influence of rater training, scale format, and rating justification are limited to the specific population of college students. These individuals are not typical of workers in full-time organizations and, as such, other environmental and social factors commonly present in organizations (e.g., the performance task,

performance appraisal experience, friendship, purpose of rating) may alter the nature of the results found here. The degree to which sample specific relationships exist within this population can only be determined following future investigations of these variables in other organizational contexts.

In addition, this study was incapable of demonstrating statistical significance among the treatment effects for convergent validity, discriminant validity, and halo due to the small sample sizes in each experimental condition. This weakness raises questions about the appropriateness of the research paradigm used here. A paradigm, in the sense employed here, is a way of addressing the phenomena in a field (Kuhn, 1970). It includes a core reasoning structure which defines the appropriate models of explanation, i.e., the ways of accounting for the phenomena of interest. Within the context of this study, the paradigm involves performance ratings and centers on the ability of three variables (i.e., training, scale format, rating justification) to influence the quality of ratings across different rater sources.

At issue here is not the soundness of the paradigm but the statistical procedures available to test it. Currently, no statistical techniques with adequate power are available to allow conclusions to be drawn from a single research study. This defect precludes definitive conclusions from being made regarding the effects that different variables may have on the quality of ratings exhibited by different rater sources. While meta-analytic techniques such as the Hedges and Olkin (1983) procedure can be used to synthesize the results of several studies which employ the same experimental

treatments, they do not possess the sensitivity necessary to demonstrate significant treatment effects within a single study. For example, in the present study a mean training ICC value of .74 (as opposed to the .243 value obtained in this study) was needed to obtain a significant training effect for convergent validity given that the no training ICCs and sample size remained constant. For discriminant validity, the mean ICC value needed for a training effect was .44 (as compared to the .349 value obtained).

Therefore, three alternatives exist for the researcher interested in advancing this line of rater source research. First, researchers can continue to extend the MTMR design to include other sources of variation, but they must realize that they are dealing with a large sample paradigm. For example, the present study needed approximately 70 ratees in each experimental condition, or a total of 560 ratees/videotapes, to achieve statistical significance for discriminant validity with the Hedges and Olkin procedure. However, it must be noted that a much larger sample would have been required to achieve statistical significance for convergent validity and halo given the ICCs found in this study. Second, a statistical procedure could be developed that is sensitive enough to detect treatment effects within a single study. Finally, researchers may need to abandon this paradigm if it is determined that the first two alternatives are not viable. A last hope would be to rely on meta-analytic techniques to determine the magnitude of treatment effects across studies and forgo conclusions based on a single study.

Conclusions

Moderate convergent validity, low discriminant validity, and a

large halo effect have dominated the rater source literature. Although many hypotheses have been advanced for the differences found among different rater sources, the research evidence is both scarce and inconsistent. The present study contributed to this body of literature by assessing the influence of rater training, scale format, and rating justification on the quality of performance ratings exhibited by self, peer, and observer raters. Prior to this study no research had systematically assessed the influence of these variables on the ratings of different rater sources. In addition, the present study used an MTMR design which allowed for an examination of the construct validity of performance ratings by the three rater sources.

In general, the data indicated that rater training, scale format, and rating justification do influence the quality of performance ratings given by different rater sources. While the results of this study were similar in some respects to those reported elsewhere, several important differences occurred. The quality of performance ratings exhibited by rater sources who did not receive training and who used the graphic rating scale was similar to that of research reported elsewhere. In contrast, rater sources who received training and used the graphic rating scale exhibited moderate convergent validity, high discriminant validity, and low halo; a combination rarely found in the literature. Rater training and the behavioral checklist apparently played a major role in improving the overall quality of performance ratings. In addition, the leniency of ratings across rater sources was also affected by the variables of interest. Specifically, training and the use of the behavioral checklist helped to reduce leniency in self-ratings in those situations where raters

had to justify their performance ratings. The practical implications of controlling lenient performance ratings in justification conditions (i.e., employee feedback) were noted.

Several areas of needed research were addressed. First, if the goal of performance measurement is to assess job performance with minimal criterion deficiency and maximum accuracy, and if rater sources are actually measuring unique aspects of a ratee's performance, then multiple rater sources are needed to increase the likelihood that all aspects of work performance are included in the appraisal process. In the present study rater sources differed in their ratings of individuals across the performance dimensions. Therefore, research must be conducted to identify which rater sources provide high-quality ratings on which performance dimensions. In addition, the present study addressed the need to examine feedback training systematically to determine what affect it may have on reducing leniency in those situations when raters must justify their ratings to the ratee. Finally, research must continue to examine the combined effects of rater training and the behavioral checklist on the quality of performance ratings provided by different rater sources. The present study documented the ability of the behavioral checklist to improve discriminant validity and reduce halo. Future research must determine what the focus of training programs should be so that convergent validity can be enhanced when the behavioral checklist is used.

Overall, this study provided valuable insights into the influence of rater training, scale format, and rating justification on the quality of performance ratings exhibited by self, peer, and observer

sources. The use of a behavioral checklist with rater training not only improved discriminant validity and reduced halo but controlled the leniency of self-evaluations that are typically exhibited by individuals who rate their own performance.

V. REFERENCES

Baird, L. (1977). Self and superior ratings of performance: As related to self-esteem and satisfaction with supervision. Academy of Management Journal, 20, 291-300.

Banks, C. G., & Murphy, K, R. (1985). Toward narrowing the research-practice gap in performance appraisals. Personnel Psychology, 38, 335-345.

Banks, C. G., & Roberson, L. (1985). Performance appraisers as test developers. Academy of Management Review, 10, 128-142.

Barrett, R. S. (1966). Influence of supervisor's requirements on ratings. Personnel Psychology, 19, 375-387.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

Bassett, G., & Meyer, H. (1968). Performance appraisal based on self-review. Personnel Psychology, 21, 421-430.

Becker, G., & Bakal, D. (1970). Subject anonymity and motivational distortion in self-report data. Journal of Clinical Psychology, 26, 207-209.

Blackburn, R., & Clark, M. (1975). An assessment of faculty performance: Some correlates between administrator, colleague, student, and self-ratings. Sociology of Education, 48, 242-256.

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

Bernardin, H. J. (1981, August). Improving rater training. Presented at the 89th annual meeting of the American Psychological Association, Los Angeles, CA.

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston, MA: Kent Publishing Co.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

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

Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. Organizational Behavior and Human Performance, 12, 105-124.

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

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C., & Dunnette, M. D. (1975). Behavior based versus trait oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

Borman, W. C., Hough, L. M., & Dunnette, M. D. (1976). Development of behaviorally based rating scales for evaluating the performance of U. S. Navy recruiters (N-TR-76-31). San Diego, CA: Navy Personnel Research and Development Center.

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

Braskamp, L. A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. American Educational Research Journal, 16, 295-306.

Burnaska, R., & Hollman, T. D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 59, 307-312.

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

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

Campbell, W. J. (1986). Construct validation of role playing exercises in an assessment center using BARS and behavioral checklist formats. Unpublished master's thesis, Old Dominion University, Norfolk, VA.

Carroll, S. J., & Schneier, C. E. (1982). Performance appraisals and review systems. Glenview, IL: Scott Foresman.

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Crooks, L. A. (1977). The selection and development of assessment center techniques. In J. L. Moses & W. C. Byham (Eds.), Applying the assessment center method (pp. 69-87). New York: Pergamon Press.

Cummings, L. L., & Schwab, D. P. (1973). Performance in organizations. Glenview, IL: Scott, Foresman and Co.

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. Academy of Management Review, 3, 635-646.

DeVries, D. L., & McCall, M. W., Jr. (1976, January). Performance appraisal: Is it tax time again? In D. L. DeVries & M. W. McCall, Jr. (Co-chairs), Managerial performance feedback: Appraisals and alternatives. Symposium conducted at the Center for Creative Leadership, Greensboro, NC.

Dickinson, T. L. (1986). Accuracy of work performance measurement: Target scores, statistics, and research issues. AFHRL technical paper, in press. Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resource Laboratory.

Dickinson, T. L., Hassett, C. E., & Tannenbaum, S. I. (1986). Work performance ratings: A meta-analysis of multitrait-multimethod studies (AFHRL-TP-86-32). Brooks AFB, TX: Training Systems Division, Brooks Air Force Human Resources Laboratory.

Dickinson, T. L., & Silverhart, T. A. (1985). Performance measurement test bed: A review of the assessment center literature. Unpublished manuscript.

Dickinson, T. L., & Silverhart, T. A. (1986, August). Training to improve the accuracy and validity of performance ratings. Paper presented at the 94th annual convention of the American Psychological Association, Washington, D. C.

Dickinson, T. L., & Tice, T. E. (1973). A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 9, 421-438.

Downy, R. G., Medland, F. F., & Yates, L. G. (1976). Evaluation of a peer rating system for predicting subsequent promotion of senior military officers. Journal of Applied Psychology, 61, 206-209.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. Annual Review of Psychology, 30, 477-525.

Dunnette, M. D., & Heneman, H. O. (1956). Influence of scale administrator on employee attitude responses. Journal of Applied Psychology, 40, 73-77.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. Personnel Psychology, 35, 105-116.

Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.

Fiske, D. W., & Cox, J. A. (1960). The consistency of ratings by peers. Journal of Applied Psychology, 44, 11-17.

French, W. L., & Bell, C. H., Jr. (1978). Organizational development: Behavioral science interventions for organizational improvement. Englewood Cliffs, NJ: Prentice-Hall.

Goldberger, A. S. (1964). Econometric theory. New York: John Wiley.

Gordon, M. E., & Petty, M. M. (1971). A note on the effectiveness of research conditions in reducing the magnitude of dissimulation on a self-report criterion. Personnel Psychology, 24, 53-61.

Griffiths, R. (1975). The accuracy and correlates of psychiatric patients' self-assessment of their work behavior. British Journal of Social and Clinical Psychology, 14, 181-189.

Guion, R. M. (1965). Personnel testing. New York: McGraw-Hill.

Gunderson, E. K., & Ryham, D. H. (1971). Convergent and discriminant validities of performance evaluations in extremely isolated groups. Personnel Psychology, 24, 715-724.

Hedge, J. W. (1982). Improving the accuracy of performance evaluations: A comparison of the methods of performance appraisal training. Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

Hedges, L. V., & Olkin, I. (1983). Regression models in research synthesis. The American Statistician, 37, 137-140.

Heneman, H. G. III. (1974). Comparison of self and superior ratings of managerial performance. Journal of Applied Psychology, 59, 638-643.

Heneman, H. G. III. (1980). Self-assessment: A critical analysis. Personnel Psychology, 33, 297-307.

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

Ilgen, D., Peterson, R., Martin, B., & Boeschen, D. (1981). Supervisor and subordinate reactions to performance appraisal sessions. Organizational Behavior and Human Performance, 28, 311-330.

Ivancevich, J. M. (1979). A longitudinal study of the effects of rater training on psychometric errors in ratings. Journal of Applied Psychology, 64, 502-508.

Jones, S. C. (1973). Self-evaluations and interpersonal evaluations: Esteem theories versus consistency theories. Psychological Bulletin, 79, 185-199.

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

Kaufman, G. G., & Johnson, J. C. (1974). Scaling peer ratings: An examination of the differential validities of positive and negative nominations. Journal of Applied Psychology, 59, 302-306.

Kavanaugh, M. J., Borman, W. C., Hedge, J. W., & Gould, R. M. (1983). Job performance measurement classification scheme for validation in the military (AFHRL-TP-85-51, AD-A164837). Brooks AFB, TX: Manpower and Personnel Division, Brooks Air Force Human Resources Laboratory.

Kavanaugh, M. J., McKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. Psychological Bulletin, 75, 34-49.

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

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored ratings scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.

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

Kraut, A. J. (1975). Prediction of managerial success by peer and training-staff ratings. Journal of Applied Psychology, 60, 14-19.

Kuhn, T. S. (1970). The structure of scientific revolution (2nd ed.). Chicago: University of Chicago Press.

Lacho, K. J., Stearns, G. K., & Villere, R. M. (1979). A study of employee appraisal systems of major cities in the United States. Public Personnel Management, 8, 111-125.

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

Latham, G. P., & Wexley, K. N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.

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

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

Lazar, R. I., & Wilkstrom, W. S. (1977). Appraising managerial performance: Current practices and future directions. New York: Conference Board.

Lewin, A. Y., & Zwany, A. (1976). Peer nominations: A model, literature critique, and a paradigm for research. Personnel Psychology, 29, 423-447.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluations of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.

Marsh, H. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. Journal of Educational Psychology, 74, 264-279.

Marsh, H. W., Overall, J. V., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. Journal of Educational Psychology, 71, 149-160.

Mascitti, A. P. (1978). Comparisons of behaviorally anchored and numerically anchored rating scales for self-, supervisory-, and peer-ratings. Unpublished doctoral dissertation, Illinois Institute of Technology, Chicago, IL.

McCall, M. W., Jr. & DeVries, D. L. (1976, September). Appraisal in context: Clashing with organizational realities. In D. L. DeVries (Chair), Performance appraisal and feedback: Flies in the ointment. Symposium conducted at the 84th annual convention of the American Psychological Association, Washington, DC.

McGregor, D. (1957). An uneasy look at performance appraisals. Harvard Business Review, 35, 89-94.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Meyer, H. (1980). Self-appraisals of job performance. Personnel Psychology, 33, 291-295.

Miner, J. D. (1968). Managerial appraisal: A capsule review and current references. Business Horizon, 11, 83-96.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. Personnel Psychology, 37, 687-702.

Nealey, & Owen, T. W. (1970). A multitrait-multimethod analysis of predictors and criteria of nursing performance. Organizational Behavior and Human Performance, 5, 345-365.

Orpen, C. (1973). An empirical assessment of high-level executives by means of a multitrait-multimethod matrix. Psychologia Africana, 15, 7-14.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. (1959). Rating scale content: III. Relationship between supervisory and self-ratings. Personnel Psychology, 12, 45-63.

Porter, L. W., Lawler, E. E., & Hackman, J. R. (1975). Behavior in organizations. New York: McGraw-Hill.

Pulakos, E. D. (1984). A comparison of training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pursell, E. D., Dossett, D. L., & Latham, G. P. (1980). Obtaining valid predictors by minimizing rating errors in the criteria. Personnel Psychology, 33, 91-96.

Regan, J. W., Gosselink, H., Hubsch, J., & Ulsh, E. (1975). Do people have inflated views of their own ability? Journal of Personality and Social Psychology, 31, 295-301.

Russell, P., & Byham, W. C. (1980). Reliability and validity of assessment in a small manufacturing company. Pittsburg: Development Dimensions International.

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

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

SAS User's Guide: Statistics. (1985). Cary, NC: SAS Institute Inc.

Schlenker, B. R. (1975). Self-presentation: Managing the impression of consistency when reality interferes with self-enhancement. Journal of Personality and Social Psychology, 32, 1030-1037.

Schneier, C., & Beatty, R. (1978). The influence of role prescriptions on the performance appraisal process. Academy of Management Journal, 21, 129-135.

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

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

Sherwood, J. J. (1966). Self-report and projective measures of achievement and affiliation. Journal of Consulting Psychology, 30, 329-334.

Shore, L. M., & Thornton, G. C. III. (1986). Effects of gender on self- and supervisor ratings. Academy of Management Journal, 29, 115-129.

Smith, D. E. (1986). Training programs for performance appraisal: A review. Academy of Management Review, 11, 22-40.

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

Sorenson, A. G. (1956). A note on the "fakability" of the Minnesota Teacher Attitude Inventory. Journal of Applied Psychology, 40, 192-194.

Stockford, L., & Bissel, H. W. (1949). Factors involved in establishing a merit rating scale. Journal of Applied Psychology, 26, 94-118.

Taylor, J. B. (1968). Rating scales as measures of clinical judgment: A method for increasing scale reliability and sensitivity. Educational and Psychological Measurement, 28, 747-766.

Teachout, M. S. (1984). The effect of anonymity, expectation of validation, and expected reward on the accuracy of self-evaluations of ability. Unpublished master's thesis, Old Dominion University, Norfolk, VA.

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

Thornton, G. C. III. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 363-371.

Thornton, G. C., & Byham, W. C. (1982). Assessment centers and managerial performance. New York: Academic Press.

Tsui, A. S. (1983). Qualities of judgmental ratings by four rater sources. Paper presented at the 91st annual convention of the American Psychological Association, Anaheim CA.

Tsui, A. S., & Ohlott, P. (1986). Multiple assessment of managerial effectiveness: Consensus in effectiveness models. Paper presented at the 94th annual convention of the American Psychological Association, Washington, D. C.

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

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. Psychological Bulletin, 72, 204-213.

Wherry, R. J. (1952). The control of bias in rating. VII: A theory of rating. Personnel Research Branch Report No. 922. Columbus, Ohio: The Ohio State University Research Foundation.

Wiley, L., & Hahn, C. (1977). Task level job performance criteria development (AFHRL-TR-75-77). Brooks AFB, TX: Brooks Air Force Human Resources Laboratory.

Williams, W., & Seiler, D. (1973). Supervisor and subordinate participation in the development of behaviorally anchored rating scales. Journal of Industrial and Organizational Psychology, 1, 1-12.

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

Zawacki, R. A., & Taylor, R. L. (1976). A view of performance appraisals from organizations using it. Personnel Journal, 55, 290-292.

Zedeck, S., & Baker, H. T. (1972). Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 7, 457-466.

Zedeck, S., & Cascio, W. (1982). Performance decision as a function of purpose of rating and training. Journal of Applied Psychology, 67, 752-758.

Zedeck, S., Imparto, N., Krausz, M., & Oleno, T. (1974). Development of behaviorally anchored rating scales as a function of organizational level. Journal of Applied Psychology, 59, 249-252.

VI. Appendix A:

Customer Role Play Instructions

Customer Role Play Instructions

Description of the Exercise

During the next 15 minutes you will be asked to participate in a role play exercise. In this exercise you and another person will each assume a role (character) and act out a real life situation. The exercise is designed to give you an opportunity to demonstrate your ability in a realistic job situation. Please behave as you would if the situation were real.

Participant's Role Instructions

It is Tuesday, 5:00 P.M. You are the manager of a Forbes' Home Improvement and Decorating center. Forbes' is a small chain of stores in the state, but has a good reputation. The store is particularly crowded. A customer has come in to the store and asked to speak to the person in charge. You walk over to speak to him.

You may handle this situation in any way you feel is appropriate. It is recommended that you act naturally as if the situation were real.

AT THIS TIME, IF YOU ARE CONFUSED ABOUT YOUR ROLE, PLEASE ASK FOR CLARIFICATION.

VII. Appendix B:

Research Study Introduction

Research Study Introduction

The study you are about to participate in is interested in various factors that influence performance appraisal ratings. We hope to learn how these factors influence the way different people rate individual performance. Unfortunately, we can not reveal these factors to you at the present time since advanced knowledge of these factors may affect the results of the study. The exact nature of the study will be explained to you in a letter that you will receive in approximately four weeks when all the data has been collected.

Today, we are going to ask you to participate in a role play exercise where you will assume the role of a store manager in dealing with an irate customer. This role play exercise will take approximately ten minutes and will be videotaped through a one-way mirror. We will then ask you to return within three weeks to participate in a three-hour performance rating session. This group session will consist of yourself and five of your peers and will involve the rating of videotaped role plays of both yourself and your peers. Do you have any questions at this time?

VIII. Appendix C:

Dimension Definitions

Dimension Definitions

PROBLEM ANALYSIS. Breaking up a problem (e.g., item or issue) into its parts such that the parts can be examined for their importance, interrelationships, or need for additional information.

PROBLEM SOLUTION. Providing actions, methods, or strategies that help in answering a problem.

PERSUASIVENESS. Attempting to influence others to an action or point of view by an overt appeal to reason or emotion, using coaxing, pleading, or arguing.

SENSITIVITY. Responding to others' feelings, needs, and points of view; letting people know you are aware of their individual situation.

IX. Appendix D:

Behavioral Checklist

Behavioral Checklist

Rater # _____ Group # _____ Subject # _____

Your Name _____

Problem Analysis

___ The manager asks the customer for more detail about the problem.

___ The manager asks the customer when the work was supposed to have been completed.

___ The manager inquires with whom the customer had dealt.

___ The manager identifies the need to check the records/contract.

___ The manager inquires whether the customer has already paid for the work contracted.

___ The manager inquires whether anyone else had access to the house.

___ The manager inquires whether the customer has proof (e.g., receipts, appraisal) of the value of the coffee table and vase.

___ The manager identifies which problems can be handled immediately and which problems require additional investigation.

___ The manager asks the customer when he wanted to have the rework done.

___ The manager asks for the customer's telephone number.

___ The manager inquires whether the house is still in the damaged condition.

Problem Solution

___ The manager establishes a time by which the customer can expect a decision.

___ The manager decides that the work will be redone if the contract matches what the customer said.

___ The manager decides to fix/repair the vase and coffee table if the customer's neighbor had no knowledge of the items being broken previously.

___ The manager establishes a time frame within which the customer will be reimbursed for the damages to the vase and coffee table.

___ The manager suggests that the coffee table may be refinished rather than replaced.

___ The manager advises the customer that the firm's insurance company will handle the problem concerning the vase and coffee table.

___ The manager advises the customer that he might be reimbursed (e.g., check, cash) for the damages at some point in the future.

___ The manager agrees to take care of everything by the following week.

___ The manager tells the customer that he will take care of the vase and coffee table but fails to specify an action plan.

___ The manager tells the customer that he didn't know what would be done to remedy the situation.

Problem Analysis

____ The manager identifies the need to talk to the employees to get their side of the story.

____ The manager inquires whether the vase and coffee table were in the same room that the work was done.

____ The manager inquires whether the customer has insurance for the vase and coffee table.

____ The manager inquires about a convenient time for him and/or his employees to see the house.

Problem Solution

____ The manager postpones his decision on all matters until he has more information.

____ The manager decides to repaint and recarpet the room.

____ The manager agrees to take care of the vase and coffee table one way or another.

____ The manager postpones a decision on the issues involving the vase and coffee table.

____ The manager advises the customer that the company is not responsible for the damages to the vase and coffee table.

Persuasiveness

___ The manager provides justifications for his inability to reach a decision that day.

___ The manager argues that it is impossible for him to make a decision without having all of the information.

___ The manager points out that there are two sides to every story.

___ The manager argues that it is necessary to talk to his employees.

___ The manager argues that they don't know what the employees will say.

___ The manager argues that they don't know that the employees damaged the vase and coffee table.

___ The manager argues that the vase and coffee table could have been ruined before the workers arrived.

___ The manager urges the customer to let him give the employees a chance to explain what happened.

___ The manager argues that the customer has to prove his case.

___ The manager urges the customer not to give him a hard time.

___ The manager provides numerous justifications for an argument.

___ The manager argues that the fact that the vase was not broken when the customer left for vacation was not proof that the employees damaged it.

Sensitivity

___ The manager is sympathetic to the customer for the problems created.

___ The manager acknowledges the legitimacy of the customer's anger.

___ The manager apologizes for the problem.

___ The manager annoys the customer by telling him that he doesn't have time to check into the matter now.

___ The manager assures the customer that he will take care of the problem personally.

___ The manager tells the customer that he is stubborn.

___ The manager loses his patience with the customer.

___ The manager assures the customer that the rework will be done to his satisfaction and asks the customer to call if there are any further problems.

___ The manager thanks the customer for bringing the matter to his attention.

___ The manager asks the customer if he is agreeable to the proposed solution.

___ The manager listens attentively to the customer.

___ The manager sympathizes with the customer's desire to have the problem corrected immediately.

___ The manager annoys the customer by telling him the store is about to close.

___ The manager assures the customer that he will work with him to get the matter resolved.

Persuasiveness

____ The manager urges the customer to let him give the employees a chance to tell their side of the story.

____ The manager attempts to convince the customer that he can't just take the customer's story.

____ The manager justifies his refusal to decide by pointing out that it was not possible to talk to the employees that day.

Sensitivity

____ The manager annoys the customer by telling him to calm down/relax.

X. Appendix E:

Graphic Rating Scale

Graphic Rating Scale

Rater #_____ Group # _____ Subject # _____

Your Name _____

We would like you to rate each individual on the five dimensions of performance defined below. Please read each definition carefully. After viewing the videotape please circle the number which you believe accurately describes the performance of the individual for that dimension.

PROBLEM ANALYSIS - Breaking up a problem (e.g., item or issue) into its parts such that the parts can be examined for their importance, interrelationships, or need for additional information.

Much Less Than Acceptable	Less Than Acceptable	Acceptable	More Than Acceptable	Much More Than Acceptable
1	2	3	4	5

PROBLEM SOLUTION - Providing actions, methods, or strategies that help in answering a problem.

Much Less Than Acceptable	Less Than Acceptable	Acceptable	More Than Acceptable	Much More Than Acceptable
1	2	3	4	5

PERSUASIVENESS - Attempting to influence others to an action or point of view by an overt appeal to reason or emotion, using coaxing, pleading, or arguing.

Much Less Than Acceptable	Less Than Acceptable	Acceptable	More Than Acceptable	Much More Than Acceptable
1	2	3	4	5

SENSITIVITY - Responding to others' feelings, needs, and points of view; letting people know you are aware of their individual situations.

Much Less Than Acceptable	Less Than Acceptable	Acceptable	More Than Acceptable	Much More Than Acceptable
1	2	3	4	5

XI. Appendix F:

Format Instructions

Format Instructions: Behavioral Checklist

The rating of individuals on the videotapes will be accomplished with a behavioral checklist. The behaviors listed on the checklist are actual behaviors displayed by individuals during the role play exercise.

HAND OUT BEHAVIORAL CHECKLIST

You will notice that there is a separate column for each dimension we just discussed. There are 15 behaviors for each dimension. The behaviors occur in an expected temporal sequence. That is, the first behavior under Problem Analysis will more than likely occur before the second behavior which will more than likely occur before the fifth and so on. However, this does not always happen.

In rating the performance of individuals on the role play exercise you are asked to check the behavior if and only if it occurs. For example, if the store manager asks the customer when the work was supposed to have been completed (a behavior under Problem Analysis) you would put a check mark next to that behavior.

The crucial thing to remember here is that you are only to check a behavior if it occurs/is observed. You are not to make inferences. For example, in Problem Solution there is a behavior: decides to repaint and recarpet. In order for you to check that behavior you need to hear the manager say: "we will repaint the walls and recarpet the floors." Or, the manager must agree to these things when the customer asks: "so are you going to repaint the walls and recarpet the floor"; and the store manager says: "yes."

Do not check the behavior if you think the manager "implied" that he would repaint and recarpet. We are only interested in actual behaviors exhibited.

Please take a few minutes now to familiarize yourself with the behaviors on the checklist. Also, compare the behaviors with the dimension definitions.

WAIT 5 MINUTES AND ASK FOR QUESTIONS

Appendix F (Continued)

Format Instructions: Graphic Rating Scale

The rating of individuals on the videctapes will be accomplished with a Graphic Rating Scale.

HAND OUT GRAPHIC RATING SCALE

The graphic rating scale consists of the four performance dimensions we just discussed: Problem Analysis, Problem Solution, Persuasiveness, and Sensitivity. Each dimension is followed by a description/definition of that dimension. Below each definition is a "numbered" scale which ranges from 1 to 5. This scale represents a continuum from ineffective to effective performance. As you can also see, each number on the scale is "anchored" by a verbal description.

After viewing a videotape, I would like you to rate the individual's performance by circling the number which you believe accurately describes the performance of the individual on that dimension. For example, after viewing the videotape you might decide that the store manager's performance on the dimension of Problem Analysis was less than acceptable. In this instance you would circle the number 2 below Problem Analysis. Please remember to rate each dimension.

I would like you to take a few minutes now to familiarize yourself with the performance dimensions and the rating scale. Please read each dimension definition carefully.

WAIT 5 MINUTES AND ASK FOR QUESTIONS

XII. Appendix G:

Training Introduction

Training Introduction

We are going to spend the next hour or so training you in how to rate the performance of individuals. We have already been over the dimensions on which performance will be evaluated and you have been introduced to the rating scale you are going to use. Although these are extremely important aspects to rating performance effectively, there are a number of other things that you should be aware of.

I want to cover 3 things which are considered essential to obtaining accurate ratings of an individual's performance. They are: CAREFUL OBSERVATION, the observation of SPECIFIC BEHAVIORS, and the need to take NOTES. Let's begin with careful observation of behavior.

1) Careful Observation of Behavior. Prior to completing the rating scale it is important that you observe carefully the task-related behaviors exhibited by the store manager. A key to obtaining accurate performance ratings is to collect as many relevant observations as possible and one way to ensure that this is done is through direct and careful observation.

2) Watch for Specific Behaviors. It would be nice to believe that the task of making specific, accurate observations can be done objectively with only minimal interference from subjective factors. Obviously, however, the subjectivity involved in evaluating people is always going to be a factor, simply because we choose to pay attention to certain things or activities while we ignore others.

It is impossible to observe everything in a given situation at the same time; while we are focusing on some attributes of a situation, we are naturally missing others. One way to use this selective attention to our advantage in terms of evaluating the performance of individuals, is to keep in mind those performance dimensions on which we are going to evaluate performance.

In our instance we are going to be rating an individuals performance on 4 dimensions: PROBLEM ANALYSIS, PROBLEM SOLUTION, PERSUASIVENESS, and SENSITIVITY. We have already been over these dimensions and their definitions. By keeping these performance dimensions in mind, they will help you to focus on those specific behaviors that are relevant.

3) Take notes. While it is not feasible to write down continually all observed behaviors, it's often beneficial to jot down behaviors as you observe them. If you don't you will have a tendency to remember especially negative behaviors, and the most recently observed behaviors. This will not give you an accurate portrayal of an individual's performance across the entire role play exercise. Therefore, it is going to be necessary to take extensive notes during the videotape so that you have an objective basis for your ratings.

In summary, there are 3 factors which are important for accurate performance ratings: observe performance carefully, watch for specific behaviors, and take notes. If you are careful in what you observe, if you focus on specific behaviors which are relevant to the performance dimensions you'll be rating, and if you take extensive notes, it should help you to be more accurate when you evaluate an individual's performance.

XIII. Appendix H:

Feedback Script

Feedback Script

No Justification Group Introduction

A skill that goes hand in hand with performance rating is the ability to give effective feedback to the performer. When an individual receives a performance rating the rating in and of itself does not help the individual's performance improve. It is necessary for the individual to be given feedback on his performance. Consequently, I want to spend a few minutes discussing exactly what makes for effective feedback skills.

Justification Group Introduction

A skill that goes hand in hand with performance rating is the ability to give effective feedback to the performer. When an individual receives a performance rating the rating in and of itself does not help the individual's performance improve. It is necessary for the individual to be given feedback on his performance. Consequently, I want to spend a few minutes discussing exactly what makes for effective feedback skills.

Remember, you will be asked to return later in the semester to participate in a feedback discussion group among yourself, your peers and several other individuals to help improve the ability of these people to rate performance accurately. Past experience has shown that face-to-face discussions are very successful for improving performance. Consequently, you will have to justify why you gave the performance ratings you did in the group discussion. Because of this it is helpful if you know a few things about giving effective feedback. Therefore, I want to spend a few minutes discussing what makes for effective feedback skills.

Feedback Lecture

Introduction

Feedback is a way of helping another person to consider changing their behavior. It is communication to a person which gives them information about some aspect of their behavior and its effect on others. As in a guided missle system, feedback helps an individual know whether their behavior is having the effect they want, it tells them whether they are "on target" as they strive to achieve their goals. For example, in our case your goal is to be able to accurately rate the performance of individuals on the videotape.

Criteria for Effective Feedback

The giving and receiving of feedback is a skill that can be acquired. When feedback is attempted at the wrong time or given in the wrong way the results will be, at best useless, and may be disastrous. Therefore developing feedback skills can be important. I want to go over some criteria that are important for effective feedback.

1) Feedback is specific rather than general. For example, it is probably more useful to learn that you "talk too much" than to have someone describe you as "dominating".

2) Feedback focuses on behavior rather than personality. It is helpful to focus on what the individual did rather than to translate their behavior into a statement about what they are. For example, the statement, "You have interrupted three people in the last half hour" is probably not something a person wants to hear, but it is likely to be more helpful than, "You are a bad-mannered oaf".

3) Feedback is well-timed. In general, feedback is most useful at the earliest opportunity after the given behavior, depending, of course, on the individual's readiness to hear it, support available from others, and so on.

4) Feedback is directed toward behavior which the individual can do something about. Frustration is increased when a person is reminded of some shortcoming over which they have no control.

5) Feedback is solicited rather than imposed. Feedback is most useful when the individual feels that they need and want it, when they have formulated the kind of question which those observing them can answer.

While these are some important criteria for giving effective feedback, it is not always easy to give feedback to others. Most of us like to give advice. Doing so suggests that we are competent and important. We get caught up in a "telling" role easily enough without testing whether our advice is appropriate to the person we are trying to help.

If the person whom we are trying to help becomes defensive, we may try to argue or pressure them. Defensiveness or denial on the part of the individual receiving feedback is a clear indication that we are going about trying to help them in the wrong way. Our timing is off or we may be simply mistaken about their behavior, but in any case, it is best to stop until we can reevaluate the situation. If we respond to the individual's resistance with more pressure, their resistance will only increase.

Feedback takes into account the needs of both the individual receiving feedback and the individual giving it. Positive feedback is welcomed by the receiver when it is genuine. If feedback incorporates the criteria given here it can become a primary means of learning about one's self.

REVIEW CRITERIA

XIV. Appendix I:

Outline for Small Group Exercise

Outline for Small Group Exercise

Introduction

I'd like to spend the first 20 minutes or so going over the role play exercise that you participated in within the last 3 weeks. To refresh your memories, I'd like to review what went on.

In this exercise you were asked to assume the role of the manager of a Forbes' Home Improvement and Decorating Center. You were told that it was 5:00pm on a Tuesday. You were further told that you would be dealing with an angry customer who had a problem. It was your task to talk to the customer and try to solve his problem. You were finally asked to pretend that you were the store manager and to deal with the individual and his problem in a way that you felt was appropriate as the store manager.

1) At this point I'd like to ask you to list 2 or 3 expectations and/or anxieties that you had just before the role play exercise started. This should only take about a minute.

WAIT FOR RATERS

Now I'd like you to share these anxieties with the rest of your peers as we put them on the board for discussion.

LIST ANXIETIES AND DISCUSS

2) Now I'd like you to each list the difficulties you encountered while dealing with the irate customer. List 2 or 3.

WAIT FOR RATERS

Once again, I'd like you to share these with the group.

LIST DIFFICULTIES AND DISCUSS

3) Finally, I'd like you to list some strategies/approaches for dealing with the customer. They can be ones you actually used or they can be strategies which you feel would be appropriate now that you have had time to think about the task.

WAIT FOR RATERS

Once again, lets discuss these strategies and see if we can come to a group consensus on which ones would be most effective.

LIST STRATEGIES AND DISCUSS

XV. Appendix J:

Pre-test and Post-test

Pre-test

We have just discussed the dimensions that you will be using to rate videotaped performances of both yourself and your peers. We are now interested in finding out what you know about performance ratings before you participate in the rest of this study. Therefore, we would like to ask you a few questions about rating performance before we proceed any further. Your answers will not be used to evaluate your performance in this study and will have no bearing on the credit you receive. It is just a way for us to establish your familiarity with this topic area. The questions should take approximately 10 minutes to complete. We ask that you give careful consideration to your responses. Please answer all questions.

RATER NUMBER _____ GROUP NUMBER _____

Part I: Matching

This section asks you to match each performance dimension we discussed with a behavioral component. For each behavioral component, choose the performance dimension that you think best represents that behavior and write the letter of that dimension in the space preceeding the behavior.

A. Problem Analysis B. Problem Solution C. Sensitivity D. Persuasiveness

Behavioral Components

____ The manager argues that they don't know that the employees damaged the coffee table and vase.

____ The manager thanks the customer for bringing the matter to his attention.

____ The manager argues that they don't know what the employees will say.

____ The manager justifies his refusal to decide by pointing out that it was not possible to talk to the employees that day.

____ The manager identifies the need to check the records/contract.

____ The manager agrees to fix/ repair the vase and coffee table if the customer's neighbor had no knowledge of the items being broken previously.

____ The manager asks the customer for more detail about the problem.

____ The manager decides that the work will be redone if the contract matches what the customer said.

____ The manager attempts to convince the customer that he can't just take the customer's story.

Behavioral Components

____ The manager assures the customer that the rework will be done to his satisfaction and asks the customer to call if there are any further problems.

____ The manager inquires with whom the customer had dealt.

____ The manager advises the customer that the firm's insurance company will handle the problem concerning the vase and coffee table.

____ The manager inquires whether the house is still in the damaged condition.

____ The manager establishes a time frame within which the customer will be reimbursed for the damages to the vase and coffee table.

____ The manager advises the customer that he might be reimbursed (e.g., cash, check) for the damages at some point in the future.

____ The manager urges the customer to let him give the employees a chance to explain what happened.

____ The manager inquires whether the customer has proof of the value (e.g., receipts, appraisal) of the coffee table and vase.

____ The manager assures the customer that he will work with him to get the matter resolved to the customer's satisfaction.

Matching Continued

A. Problem Analysis B. Problem Solution C. Sensitivity D. Persuasiveness

Behavioral Components

____ The manager loses his patience with the customer.

____ The manager argues that the vase and coffee table could have been ruined before the workers arrived.

____ The manager establishes a time by which the customer can expect a decision.

____ The manager inquires whether the customer has already paid for the work contracted.

____ The manager advises the customer that the firm is not responsible for the damages to the vase and coffee table.

____ The manager argues that it is necessary to talk to his employees.

Behavioral Components

____ The manager tells the customer that he is stubborn.

____ The manager argues that it is impossible for him to make a decision without having all of the information.

____ The manager agrees to take care of everything by the following week.

____ The manager identifies which problems can be handled immediately and which problems require additional investigation.

____ The manager asks the customer if he is agreeable to the proposed solution.

____ The manager sympathizes with the customer's desire to have the problem corrected immediately.

Part II: Short Answers

1. If you were responsible for observing and then rating an individual's performance, what are some of the things you would do to make sure your rating was accurate?

2. People often receive performance feedback from their supervisor in a formal performance appraisal feedback interview. What do you believe are some important components of an effective feedback discussion?

Post-test

You have just completed rating the performance of several individuals on the role play exercise. We are now interested in finding out what you have learned about performance ratings from this study. Therefore, we would like to ask you a few questions about rating performance before you leave. Once again, your answers will not be used to evaluate your performance in this study and will have no bearing on the credit you receive. It is just a way for us to establish what you have learned about this topic area. The questions should take approximately 10 minutes to complete. We ask that you give careful consideration to your responses. Please answer all questions.

RATER NUMBER _____ GROUP NUMBER _____

Part I: Matching

This section asks you to match each performance dimension we discussed with a behavioral component. For each behavioral component, choose the performance dimension that you think best represents that behavior and write the letter of that dimension in the space preceeding the behavior.

A. Problem Analysis B. Problem Solution C. Sensitivity D. Persuasiveness

Behavioral Components

____ The manager argues that the fact that the vase was not broken when the customer left for vacation was not proof that the employees damaged it.

____ The manager agrees to take care of the vase and coffee table one way or another.

____ The manager annoys the customer by telling him that the store is about to close.

____ The manager argues that the customer has to prove his case.

____ The manager decides to repaint and recarpet the room.

____ The manager postpones his decision on all matters until he has more information.

____ The manager asks for the customer's telephone number.

____ The manager postpones a decision on the issues involving the vase and coffee table.

____ The manager acknowledges the legitimacy of the customer's anger.

____ The manager inquires whether anyone else had access to the house.

Behavioral Components

____ The manager inquires when the work was supposed to have been completed.

____ The manager tells the customer that he didn't know what would be done to remedy the situation.

____ The manager is sympathetic to the customer for the problems created.

____ The manager assures the customer that he will take care of the problem personally.

____ The manager provides numerous justifications for an argument.

____ The manager inquires about a convenient time for him and/or his employees to see the house.

____ The manager inquires when the customer wanted to have the rework done.

____ The manager inquires whether the customer has insurance for the vase and coffee table.

____ The manager inquires whether the vase and coffee table were in the same room that the work was done.

____ The manager suggests that the coffee table may be refinished rather than replaced.

Matching Continued

A. Problem Analysis B. Problem Solution C. Sensitivity D. Persuasiveness

Behavioral Components

____ The manager annoys the customer by telling him that he doesn't have time to check into the matter now.

____ The manager annoys the customer by telling him to calm down.

____ The manager apologizes for the problem.

____ The manager tells the customer that he will take care of the coffee table and vase but fails to specify an action plan.

____ The manager identifies the need to talk to the employees to get their side of the story.

Behavioral Components

____ The manager points out that there are two sides to every story.

____ The manager urges the customer to give the employees a chance to tell their side of the story.

____ The manager urges the customer not to give him a hard time.

____ The manager provides justifications for his inability to reach a decision that day.

____ The manager listens attentively to the customer.

Part II: Short Answers

1. If you were responsible for observing and then rating an individual's performance, what are some of the things you would do to make sure your rating was accurate?

2. People often receive performance feedback from their supervisor in a formal performance appraisal feedback interview. What do you believe are some important components of an effective feedback discussion?

XVI. Appendix K:

Post-Experimental Questionnaire

Post-Experimental Questionnaire

1. RATER #: _____ GROUP #: _____

2. Sex: Male Female (Circle one)

3. Age: _____

4. Ethnic Origin: White Black Hispanic Asian Other (Circle one)

5. Class Rank: Freshman Sophomore Junior Senior (Circle one)

6. Would you be interested in participating in another research study similar to this one?

   Yes No (Circle one)

7. Will the experimenter be able to match your name to the performance ratings you gave? (Circle a number)

Not at all		Quite a	To a great	
	Somewhat	bit	extent	Completely
1	2	3	4	5

8. How confident were you in assessing an individual's performance? (Circle a number)

Not at all	Somewhat		Quite	Extremely
Confident	Confident	Confident	Confident	Confident
1	2	3	4	5

9. Was the experiment a learning experience for you? (Circle a number)

Not at all		Quite a	To a great	
	Somewhat	bit	extent	Completely
1	2	3	4	5

10. How confident are you that your ratings are accurate measures of an individual's performance? (Circle a number)

Not at all	Somewhat		Quite	Extremely
Confident	Confident	Confident	Confident	Confident
1	2	3	4	5

11. Will you be held accountable for the performance ratings you gave? Yes____ No____ (Check one) If yes, how will you be held accountable.

_____

_____

12. Can you be identified with the performance ratings you gave in this experiment? (Circle a number)

Not at all		Quite a	To a great	
	Somewhat	bit	extent	Completely
1	2	3	4	5

13. Did the rating scale you used enable you to adequately document an individual's performance? (Circle a number)

Not at all		Quite a	To a great	
	Somewhat	bit	extent	Completely
1	2	3	4	5

14. Will this experiment enhance Old Dominion's image? (Circle a number)

Not at all		Quite a	To a great	
	Somewhat	bit	extent	Completely
1	2	3	4	5

15. Based on the rating scale you used, how confident are you that your ratings accurately reflect the performance of those individuals you rated? (Circle a number)

Not at all Confident	Somewhat Confident	Confident	Quite Confident	Extremely Confident
1	2	3	4	5

16. Will the individuals you rated on the videotapes know what performance ratings you gave them?
Yes_____ No_____ (Check one). If yes, how will they know?

_____

_____

17. Were the instructions for the rating form you used clear and easy to understand? (Circle a number)

Not at all		Quite a	To a great	
	Somewhat	bit	extent	Completely
1	2	3	4	5

18. The experimenter will use the data from this study:
(Circle a letter)

A) for psychological research on performance ratings only.

B) to evaluate the performance of the individuals who participated in the role play exercise.

C) in a feedback discussion group to help improve the ability of individuals to rate performance effectively.

19. Do you think this research contributes to society?
(Circle a number)

Not at all			Quite a bit	To a great extent	Completely
1	Somewhat 2		3	4	5

Biographical Statement

The author was born in Centereach, New York on February 19, 1959. He received his B.A. degree from Southeastern Massachusetts University in June 1981 and his M.S. degree from Old Dominion University in August, 1984. Journal publications include: Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, Jr., R. L. (1986). Influence of assessment center methods on assessor ratings. Personnel Psychology, 39, 567-578. During his graduate training at Old Dominion University the author held teaching assistantship positions in Experimental Methods and Social Psychology. He also served as an instructor for Quantitative Methods (Statistics). The author is a member of the American Psychological Association, the Academy of Management, and the Southeastern Psychological Association.