Summer 1988

Performance Appraisal Ratings as a Function of Source of Ratings and Purpose of the Appraisal

Richard J. Tannenbaum
Old Dominion University

Performance Appraisal Ratings as a Function of Source of Ratings and Purpose of the Appraisal

by

Richard J. Tannenbaum
B.A. May 1981, State University of New York at Stonybrook
M.S. August 1983, Rensselaer Polytechnic Institute

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
August, 1988

Approved by:

Terry L. Dickinson (Director)

DEDICATION

To the eternal spirit of Poppy Al, for the love you always shared with each one of us, and the greater meaning you brought to all our lives.

ACKNOWLEDGEMENTS

Without the support of many people the road I have travelled as a graduate student surely would have been a much rockier one.

My family has been a great source of strength. My parents have always been there for me, and for that I can not thank them enough. Your love touches me wherever I go.

I extend my gratitude to my dissertation committee, Drs.: Terry Dickinson, Glynn Coates, Albert Glickman, and Diane Yarosz for their support and guidance throughout this arduous task.

I would especially like to thank Dr. Terry Dickinson, who chaired my dissertation committee, for the patience, understanding, and wisdom that he shared with me throughout my tenure as a graduate student. Also, thank you Dr. Coates for always being available to me to answer my statistical questions, even when some of them where clearly from "left field". In addition, thank you Dr. Anthony "Skip" Dalessio, you were a great positive influence on me, and I am truly fortunate to know you.

Thank you Steve, Todd, and Eric for everything, but especially for being great friends. If I had to do it all over again, I would have, if only for the lifetime friendships I have made with all of you.

Finally, thank you Lor, I am a much more caring and sensitive person because of you, and I will always cherish the times we spent together.

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (concluded)

List of Tables

List of Tables (concluded)

Performance Appraisal Ratings as a Function of Source of Ratings and Purpose of the Appraisal

I. INTRODUCTION

Performance appraisal is a systematic procedure for obtaining job performance information. The results of performance appraisal are used as inputs into many important organizational decisions such as promotion, compensation, employee development, and determination of training needs (Latham & Wexley, 1981). Performance appraisal procedures are used in a wide variety of organizations, both industrial and nonindustrial (French, 1982). In support of the pervasiveness of performance appraisal, Locher and Teel (1977) reported that over 90% of the organizations included in their study utilized such a procedure.

Measures of job performance are classified into two basic categories: (a) objective and (b) subjective (Cascio, 1982). Objective measures are further divided into production data and personnel data (Guion, 1965). Production data include measures of quantity of units produced and dollar volume of sales. These are direct measures of production. Indirect measures of production include learning time and commissions earned (Guion, 1965). Personnel data are independent of specific jobs, and include measures of tenure, absences, rate of advancement, and accidents (Guion, 1965). While objective measures of job performance are intuitively appealing because of their highly quantifiable nature, they are subject to such measurement deficiencies as unreliability and contextual constraints (Cascio, 1982).

Performance unreliability refers to a lack of consistency in a job

performance measure. Rothe (1978) concluded that production output for an individual was highly unreliable and inconsistent over time. Guion (1965) noted that in one of the studies conducted by Rothe (not cited), over a 38 week period, intra-individual productivity correlations ranged from a low of 0.03 to a high of 0.91. Thus, obtaining an accurate indication of job performance is quite difficult with an objective measure.

Another shortcoming of objective measures of job performance is that often an individual's level of job performance is affected by factors not within the individual's control (Guion, 1965). For example, an assembly-line worker's productivity is influenced by the pace of the conveyer belt, which may or may not be under the individual's control (Guion, 1965). Similarly, a salesperson's sales volume is influenced by the particular territory. In addition, for some jobs such as middle management positions, there does not appear to be objective measures of job performance (Cascio, 1982). Landy (1985) also indicated that measures of job performance in personnel data are typically present in less than 5% of the cases examined, rendering personnel data virtually useless for performance appraisals.

Confronted with the weaknesses present in the objective measures of job performance, much attention has focused upon subjective measures. The most common subjective measure utilized is performance ratings. Guion (1965) reported that of the validation studies published in the Journal of Applied Psychology and Personnel Psychology between 1950 and 1955, 81% percent used ratings as criteria. Landy and

Trumbo (1980) extended this review by surveying validation studies in the Journal of Applied Psychology from 1965 to 1975 and discovered that 72% percent of the studies used ratings as the primary criterion. Although performance ratings are the most commonly used performance measure, they too are subject to several deficiencies, such as leniency, severity, halo, and central tendency errors (Smith, 1986). These rating errors affect both the validity and the accuracy of performance ratings (DeCotiis & Petit, 1978; Saal, Downey, & Lahey, 1980).

In response to the limiting effects that rating errors have on accuracy and validity estimates, a great deal of research has been conducted to improve the psychometric qualities of performance ratings. Research has focused on the rating instrument itself as a means to reduce rating errors (e.g., Borman, 1979; Borman & Vallon, 1974; Dickinson & Zellinger, 1980). Comparisons have been made between various rating scales in terms of leniency, discriminability, halo, user acceptance, and reliability. For example, behaviorally anchored rating scales have been compared to graphic rating scales (Burnaska & Hollmann, 1974; Campbell, Dunnette, Arvey, & Hellervik, 1973). Berkshire and Highland (1953) compared forced-choice rating scales to graphic rating scales. Mixed standard scales have been compared to behaviorally anchored rating scales (Dickinson & Zellinger, 1980; Saal & Landy, 1977). In addition, a recent meta-analysis of multitrait-multimethod studies of work performance ratings (Dickinson, Hassett, & Tannenbaum, 1986) concluded that the use of behaviorally-oriented

rating scales yielded higher quality ratings (i.e., greater convergent validity and/or lower method bias) than did graphic rating scales. Discriminant validity was increased through the use of rating scales requiring several ratings per performance dimension.

McIntyre, Smith, and Hassett (1984) suggested that attempts to improve the psychometric qualities of performance ratings would benefit by focusing on process-related factors. The conceptualizing of performance appraisal from a more dynamic perspective is evident by the large body of research which has accumulated investigating the effects of rater training on reducing rating errors and increasing rating accuracy (e.g., Bernardin & Pence, 1980; Borman, 1975; 1979; Latham, Wexley, & Pursell, 1975; Woods, 1987; Zedeck & Cascio, 1982).

One training approach is rater error training. In this training strategy, attempts are made to reduce the various rating errors by providing the raters with examples of the more common errors, thereby increasing the rater's awareness. Once the raters are familiar with the types of rating errors they are encouraged to avoid making them (Smith, 1986).

Another training approach has been referred to as performance dimension training (Smith, 1986). This procedure consists of familiarizing the raters with the performance dimensions. This is accomplished by providing the raters with job specifications, and/or having the raters participate in the development of the rating scale (Smith, 1986).

The last common approach to rater training has been labeled frame-

of-reference training. As depicted by Bernardin and Beatty (1984), frame-of-reference training involves presenting raters with normative standards. These standards serve as true scores of ratee performance. The goal is to instill in the raters a common perception of performance standards, comparable to the presented standards, so that ratees' performance is assessed similarly by the different raters (McIntyre et al., 1984). The effectiveness of the various training strategies has been reviewed elsewhere (see Smith, 1986; Spool, 1978).

Models of Performance Appraisal

Several models of performance appraisal have been developed proposing how many of the researched appraisal factors interact to influence performance ratings. These models serve as a useful framework for understanding performance appraisal and for providing us with avenues to explore for future research. Three such models will subsequently be discussed. The salient points of each model will be briefly introduced and then integrated to define an area of performance appraisal that has been neglected by past research.

DeCotiis and Petit (1978) developed a model of performance appraisal based upon Taft's theory of interpersonal judgment. Their basic proposition stated that accurate performance ratings occurred when the rater was motivated to rate accurately, used relevant performance standards, and had the appropriate opportunity to rate the ratee.

Of particular interest in this model was the emphasis placed upon the perceived consequences of the rating task as it affected rater

motivation. DeCotiis and Petit (1978) noted that there was a general reluctance by the rater to complete the rating form accurately. This resistance was attributed to the relationship which existed between the purpose of the appraisal, and its consequences for the rater and ratee. When the appraisal was conducted for research purposes it posed no threat to either the rater or ratee. This was because the outcome of the appraisal did not affect any organizational rewards or punishments (DeCotiis & Petit, 1978). Under these conditions the ratings obtained tended to be an accurate reflection of ratee behavior. In contrast, when the appraisal was conducted for administrative purposes, and consequently affected organizational rewards and punishments, both the rater and ratee perceived performance appraisal as a negative experience. Particularly, the rater felt uncomfortable in providing any negative feedback to the ratee, and attempted to avoid this situation by assigning inaccurate, inflated ratings (DeCotiis & Petit, 1978).

This model thus suggests that it may be inappropriate to make general conclusions concerning the quality of performance appraisal ratings without considering the impact of the perceived consequences of those ratings on both the rater and the ratee.

In 1980 Landy and Farr presented a process model of performance ratings which represented a compilation of the research conducted spanning a 30 year time period. Based upon their review of the research literature, Landy and Farr (1980) conceptualized their model as consisting of several interrelated systems. The context component

of the model was comprised of position characteristics, organizational characteristics, and the purpose of the appraisal. The rating process component of the model was divided into the cognitive process of the rater and the administrative process of the organization (Landy & Farr, 1980).

In common with DeCotiis and Petit (1978), Landy and Farr (1980) emphasized the importance of both the rater and the purpose of the appraisal. The type of rater, or source of ratings, was considered to be an important rater characteristic. This factor was indirectly addressed by DeCotiis and Petit (1978), when they briefly discussed the general lack of ratings agreement across different levels of the organizational hierarchy. However, Landy and Farr (1980) directly recognized this rater characteristic by including a partial review of the literature comparing the different types of potential rating sources (self-ratings, supervisor ratings, and peer ratings), in terms of the psychometric quality of performance ratings. Their review supported the conclusion of DeCotiis and Petit (1978) that there was a lack of agreement among the different sources of rating. Particularly, self-ratings were the most discrepant and tended to be characterized by greater leniency and less halo errors than the other sources. This topic of discussion will be addressed in greater detail at a later point in this text. It will suffice at this point merely to indicate that who shall rate is an important issue in understanding performance ratings.

Landy and Farr (1980) similarly concluded that ratings obtained

under administrative conditions were more lenient than those obtained under research conditions. Furthermore, Landy and Farr (1980) stated that the purpose of the appraisal substantially impacted upon the cognitive process of the rater.

It would also seem justified to hypothesize, based upon the models of DeCotiis and Petit (1978) and Landy and Farr (1980), that the various purposes for which the appraisal is conducted may represent different perceived consequences for different types of raters. For example, it is plausible to assume that a supervisor's ratings of a subordinate for promotion, in that supervisor's department, would be perceived, by that rater, as more consequential than if the ratings were conducted for employee retention purposes (DeNisi, Cafferty, & Meglino, 1984). However, it is equally plausible to assume that an incumbent's self-ratings for employment retention would be more consequential, for that rater, than if the self-ratings were conducted for promotion purposes. Thus, the performance ratings assigned to a ratee would be affected by the purpose of the appraisal and the source of the ratings.

Further recognizing the importance of the cognitive operations of the rater in performance ratings, DeNisi et al. (1984) developed a cognitive model of performance appraisal. This model is similar to the DeCotiis and Petit (1978) and Landy and Farr (1980) models in that the emphasis is placed upon the rater and the purpose of the appraisal. However, this model also distinguishes itself by proposing that the purpose of the appraisal plays a more involved role in performance

ratings than considered previously.

Specifically, in addition to establishing the reason for collecting performance information, the purpose also determines the kinds of performance information sought by the rater (DeNisi et al., 1984). Appraisals conducted to provide feedback cause raters to seek out goal-oriented behaviors, while appraisals conducted for administrative decisions cause raters to seek out trait-oriented behaviors. In addition, the ratings supplied by a rater vary depending upon the purpose of the appraisal. This is due to the rater's belief about the impact of his/her rating decision (DeNisi et al., 1984). Once again, it seems reasonable to propose that different sources of rating will not have similar perceptions of their ratings, as defined by the purpose of the appraisal. Thus, there should be a purpose of the appraisal by source of the ratings interaction.

Each of the models presented above functioned as a framework, helping to integrate existing research findings and suggest new areas of potential study. One area of potential study suggested by these three models is the interactive role of the purpose of the appraisal and the source of the ratings. Although each of these factors was introduced as contributing to the variance in performance appraisal ratings, none of the models formally proposed the interaction between these two factors. However, the experimenter believes that enough theoretical justification has been presented to warrant an empirical investigation of the interaction between the two factors. Therefore, the purpose of this research study was to examine the joint influence

of the purpose of the appraisal and the source of the ratings on the quality of performance appraisal ratings.

Based upon the above, a formal review of the research literatures concerning the purpose of the appraisal and source of performance ratings will be presented. It will be demonstrated that despite this potential interaction, each research base has largely neglected the other.

Important Characteristics of Studies Investigating Appraisal Purpose

Explicit Appraisal Purposes. A common characteristic of the studies investigating the effects of appraisal purpose has been an almost exclusive focus upon explicit rather than implicit purposes (e.g., Borresen, 1967; Centra, 1976; Murphy, Balzer, Kellam, & Armstrong, 1984; Sharon & Bartlett, 1969; Zedeck & Cascio, 1982). Explicit purposes are operationalized by varying instructional sets to different raters. Each purpose defines an intended use of the ratings. A correct understanding of the instructional set by the raters constitutes a successful purpose manipulation. The raters are expected to rate performance only within the context of the overtly defined purpose.

Appraisal Purposes Within Explicit and Implicit Research Contexts. The vast majority of studies which have examined the role of appraisal purpose can be viewed to have taken place within either an explicit or implicit research context. An explicit research context is defined, by this researcher, as a situation where either one or both of the following study characteristics apply: (a) the raters are

instructed to rate the job performance of hypothetical persons depicted in vignettes (e.g., Williams, DeNisi, Blencoe, & Cafferty, 1985; Zedeck & Cascio, 1982); (b) the raters are instructed to role-play and pretend that their ratings are actually going to affect the defined purpose.

An implicit research context is defined, by this researcher, as a situation where either one or both of the following study characteristics apply: (a) the intended use of the ratings is not typically encountered by the particular rater group; (b) there exists an incongruity between the the task itself and the context surrounding the task. The presence of either of these characteristics increases the artificiality of the experimental setting, and may produce a reactive arrangements effect (Campbell & Stanley, 1963). Examples of the first condition are studies such as Centra (1976) and Driscoll and Goodwin (1979). In both cases, students were instructed to rate the job performance of their professor for making decisions about salary, tenure and promotion. Although students do typically evaluate their professor's performance, it is not typically the case that such ratings are used for these purposes. Furthermore, in these types of studies, only a relatively small number of classes participate. However, students (raters) are frequently members of more than one class. This multiple-membership makes it highly probable that the same students who evaluated one professor for a salary, tenure, and promotion purpose, will not do the same for another professor. This "treatment contamination" increases the artificial nature of the task and suggests that the ratings may have actually been done for some unspecified

research project.

Examples of the second condition, signifying an implicit research context, are studies such as McIntyre, Smith, and Hassett (1984) and Murphy, Balzer, Kellam, and Armstrong (1984). Although, in these studies, the purposes included may have been more plausible for student raters (e.g., decisions affecting the hiring of teaching assistants), the students received course and/or research credit for their participation. The incongruity between rating performance for a defined "real" purpose and simultaneously receiving research credit for doing so, again suggests that the ratings may have actually been done for some unspecified research project.

The psychometric results obtained from these experimental designs are comparable to those obtained from studies not defined by research contexts (e.g., Bernardin, Orban, & Carlyle, 1981). However, the effect sizes for the studies using vignettes are greater (Murphy, Herr, Lockhart, & Maguire, 1986). The directional similarity of the results obtained across these experimental designs permits cross-study comparisons.

Purpose of the Appraisal: A Review of the Literature

Several studies have examined the effects of the purpose of the appraisal (e.g., research purposes compared to administrative purposes) on rating accuracy and the psychometric qualities of performance ratings (e.g., Aleamoni & Hexner, 1980; Centra, 1976; Driscoll & Goodwin, 1979; Gmelch & Glasman, 1977; Meier & Feldhusen, 1979; Murphy, Balzer, Kellam, & Armstrong, 1984). However, few studies have examined

how the purpose of the appraisal interacts with other factors involved in performance appraisal. Studies have investigated the interactive effects of purpose of the appraisal and rater training (e.g., McIntyre, Smith, & Hassett, 1984; Warmke & Billings, 1979; Zedeck & Cascio, 1982), rater trust and rater characteristics (e.g., Bernardin, Orban, & Carlyle, 1981), absolute and relative decision outcomes (e.g., Williams, DeNisi, Blencoe, & Cafferty, 1985), and expectation of self-rating validation (e.g., Farh & Werbel, 1986). The relevant characteristics of these latter studies are summarized in Appendix A, Table A-1.

Warmke and Billings (1979) examined the effects of four different training conditions on the quality of performance ratings. The four training conditions were: (a) a lecture on rating errors (halo, leniency, central tendency, and similar-to-me effects), (b) discussion about rating errors, (c) participation in scale development, and (d) no training. The quality of performance ratings was measured by the extent of halo, leniency, and variability present in the ratings. The participants were head nurses and assistant head nurses at a university hospital who rated staff nurses for both experimental and administrative purposes. The ratings were made using two different graphic rating scales, measuring five and nine dimensions of performance, respectively. For the experimental purpose, half of the ratings were made during the first 2 weeks of the study following training (order 1), and the other half were made during the last 2 weeks of the study (order 2). For the administrative purpose, the

ratings were obtained from personnel files, 2 months after the training (Warmke & Billings, 1979).

The results indicated that for the ratings obtained for the experimental purpose and completed within a week after training (order 1), the lecture training and the scale construction groups were superior to the other training conditions in increasing variability. The scale construction group was also superior to the other conditions in controlling for halo. However, leniency was not affected by any training condition. No training effect was found for those ratings obtained during the last 2 weeks of the study (order 2) (Warmke & Billings, 1979).

Similarly, no training effect was found for those ratings obtained for the administrative purpose. The analysis did reveal that greater halo was present in the ratings of the administrative purpose than in the experimental purpose (Warmke & Billings, 1979).

The interactive effects of rater training and purpose of the appraisal on rating accuracy and discriminability were examined by Zedeck and Cascio (1982). The participants in this study were undergraduate psychology and business students. Rater training consisted of presentation and examples of common rating errors (i.e., leniency, halo, central tendency, first impressions) and outside readings concerning rater training and performance appraisal. In addition, rating practice sessions and feedback were provided to the participants, along with role-play sessions. Another group of raters received no training during this same time period and served as a

control group (Zedeck & Cascio, 1982).

Raters were randomly assigned to one of three purpose conditions: (a) recommendation for employee development, (b) awarding a merit raise, or (c) retaining a probationary employee. Ratings were made on five performance dimensions. These dimensions were presented to the raters in 33 vignettes. Each vignette described the performance of the target person, a supermarket checker, on each dimension. The dependent measure was the standard deviation of the ratings within raters across the 33 vignettes.

The results revealed that only a purpose main effect was significant. Specifically, those participants who made ratings for the merit raise condition displayed less variability in their ratings than did the other groups (Zedeck & Cascio, 1982). In addition, the results indicated that rater strategy varied with the purpose of the ratings. Raters weighted, combined, and integrated identical dimensions of performance differently depending upon the purpose of the appraisal (Zedeck & Cascio, 1982).

McIntyre et al. (1984) similarly examined the effects of rater training (rater error training, frame-of-reference training, both rater error and frame-of-reference training, and no training) and appraisal purpose (hiring, feedback, and research) on rating accuracy. The participants in this study were undergraduate students. The rating stimuli consisted of four videotaped lectures. Ratings were made across 12 performance items. Two of the dependent measures consisted of assessments of halo and leniency (McIntyre et al., 1984).

The results revealed a significant purpose main effect for leniency. Ratings in the research only condition were less lenient than in the feedback and the hiring conditions (McIntyre et al., 1984). There was no significant training by purpose interaction, a result also obtained by Zedeck and Cascio (1982). Appraisal purpose did not appear to affect halo. In contrast, there was a significant training effect. The frame-of-reference training condition was significantly closer to the "true halo" (expert raters' halo) than were the other training conditions (McIntyre et al., 1984).

The relationships between the purpose of the appraisal, rater trust, and rater characteristics were examined by Bernardin et al. (1981). Two police departments supplied performance appraisal ratings for different purposes. One purpose was for feedback only and the other was for a promotion decision. The effect of appraisal purpose on leniency was of primary interest. The raters consisted of police department sergeants; the ratees consisted of rookie patrol officers. The ratings were made on 11 performance items measured by a 9-point graphic rating scale (Bernardin et al., 1981).

The results indicated a significant purpose main effect. Significantly greater ratings were obtained for promotion purposes than for feedback purposes. In addition, raters expressing greater trust in the appraisal system displayed less leniency in their ratings than did raters expressing lower trust (Bernardin et al., 1981). Further analysis indicated a significant cognitive complexity main effect. However, contrary to expectations, there was no significant

interaction between rater trust and appraisal purpose or between cognitive complexity and purpose of the appraisal (Bernardin et al., 1981).

In a two-part study, Williams et al. (1985) investigated the effects of appraisal purpose and outcome decisions on both performance information integration and acquisition. Experiment I (reviewed here) was designed to determine if performance information was used differently, leading to varied ratings, for different appraisal purposes. Participants were undergraduates who were provided with vignettes of performance concerning a budget preparation task (Williams et al., 1985).

Ratings were made for one of three purposes: (a) a salary increase, (b) a promotion recommendation, or (c) a remedial training referral. In addition, the ratings were made for two outcome decisions: (a) a relative decision, necessitating the comparison among the target individuals depicted in the vignettes, or (b) an absolute decision, not requiring any comparison among the target individuals. Ratings were made using a 7-point Likert-type rating scale (Williams et al., 1985).

The results revealed a significant main effect for both outcome decision and appraisal purpose. The ratings of the absolute outcome group were significantly greater than those of the relative outcome decision group. The ratings of the remedial training condition were significantly greater than those of the promotion condition and the salary increase condition. However, no significant outcome decision by

appraisal purpose interaction was obtained (Williams et al., 1985).

Farh and Werbel (1986) examined the effects of appraisal purpose and the expectation of ratings validation on the leniency in students' self-ratings of their level of class participation. It was hypothesized that students' self-ratings conducted for administrative purposes (course grade) would be more lenient than self-ratings conducted for research purposes. It was also hypothesized that when the self-ratings were conducted under a condition of high expectation of validation (self-ratings compared with an independent measure of participation), the self-ratings would be less lenient than when conducted under a condition of low expectation of validation (Farh & Werbel, 1986).

The results revealed significant main effects for both appraisal purpose and expectation of validation. Participants in the grading purpose condition displayed greater leniency in their self-ratings than did those in the research purpose condition. The participants in the low expectation of validation condition similarly had greater leniency in their self-ratings than did those in the high expectation of validation condition. The appraisal purpose by expectation of validation interaction was not significant. Moreover, significantly less variable ratings occurred under conditions of greatest leniency (Farh & Werbel, 1986).

Summary of Main Findings Regarding Purpose of the Appraisal

Purpose of the appraisal has been operationalized in many different ways in the literature. The most frequently used definitions

of purpose have been merit pay, promotion, feedback, training, and research. Typically the comparison has been between some administrative purpose and a research only purpose in terms of the psychometric qualities of the obtained ratings.

Leniency error, a tendency to assign a higher rating to an individual than is justified by the behavior of that individual, has been the most researched of the rating errors. The majority of the research studies concluded that greater leniency occurred under administrative purposes than under research only purposes.

Variability error, a failure to differentiate between ratees within a dimension, and halo error, a failure to differentiate between rating dimensions within a ratee, have received much less attention. Thus, conclusions drawn from this limited research base must be considered tentative. In general, however, less variable ratings have been obtained under administrative purposes than under research purposes. In contrast, greater halo has been reported in ratings obtained under administrative conditions than under research only conditions. However, this latter result was based upon the findings of a single study, illustrating the paucity of research that has been conducted examining the impact of appraisal purpose on halo.

In summation, Landy and Farr (1983) best described the state of the research concerning the purpose of the appraisal by concluding that too little information was currently available to draw firm conclusions about the impact of appraisal purpose on ratings. "The intuitive importance of purpose, especially perhaps of perceived purpose, demands

more research effort in this area" (p. 153).

Sources of Performance Ratings

It was proposed that there exists a potential appraisal purpose by source of ratings interaction. However, it is evident from the above review, that studies examining the interactive role of the purpose of the appraisal have neglected the variance accounted for by the source of the ratings. Typically only one type of rater, either a supervisor, subordinate, or incumbent was considered.

The importance of considering multiple sources of ratings has been addressed by Lawler (1967). Obtaining ratings from various sources, such as the supervisor, peer, and incumbent, clarifies the perceptions of each member and positively affects motivation (Lawler, 1967). In addition, decision quality can be improved by using multiple raters, due to the unique perspective each rater may have in terms of the target individual's job performance (Lawler, 1967). This will increase the probability of obtaining a more complete description of the target individual's total contribution to the organization (Latham & Wexley, 1981). Moreover, greater accuracy has been attributed to multiple rating systems than to rating systems involving only a single rating source (Bernardin & Beatty, 1984).

Latham and Wexley (1981) presented evidence from a number of organizations describing the percentage of the different rating sources used by each organization. The two most widely used sources of ratings were the immediate supervisor (approximately 90% of the organizations), and the incumbent (approximately 10% of the organizations). While it

is clear that most organizations prefer to have the immediate supervisor perform the ratings, it is important that the incumbent's self-ratings are taken into consideration.

Most performance appraisal interviews involve the immediate supervisor providing feedback to the incumbent concerning the incumbent's job performance strengths and weaknesses. This places the supervisor in the role of judge and places the incumbent in a defensive role. Often this results in incumbents denying their weaknesses, and decreases incumbents' motivation to improve their subsequent job performance (Kay, Meyer, & French, 1965). The greater the disparity between the incumbent's self-ratings and the supervisor's ratings of the incumbent's job performance, the lower will be the incumbent's level of satisfaction, motivation, and job effectiveness (Bernardin & Abbott, 1985).

It is therefore important to evaluate the relationship between these two sources of appraisal ratings. By understanding how each source perceives job requirements and job performance, areas of disagreement can be identified and addressed (Bassett & Meyer, 1968; Hobson, Mendel, & Gibson, 1981). This should result in more effective communication during the appraisal interview and more positive outcomes following the appraisal interview. The literature focusing predominantly on the relationship between supervisor ratings and incumbent self-ratings is summarized in Appendix A, Table A-2.

One of the earliest studies comparing different rating sources was conducted by Parker, Taylor, Barrett, and Martens (1959). The study

was conducted to determine the effect of the amount of information supplied on the rating format on the rater's subsequent ratings. Three separate ratings were obtained on each member of a group of clerical employees. One rating was supplied by the immediate supervisor. The second rating was supplied by the second-level supervisor. The third rating was a self-rating supplied by the clerical employee. All ratings were used only for research purposes. The ratings were made across eight performance items, using a graphic rating scale. For each rater, estimates of leniency and halo were assessed (Parker et al., 1959).

The results indicated that there existed large disagreements between self-ratings and supervisor ratings. The results also indicated that both rating sources displayed leniency in their ratings. However, greater leniency was evident in the self-ratings than in the supervisor ratings. In contrast, less halo was present in the self-ratings. The self-ratings also displayed less variance than the supervisor ratings (Parker et al., 1959).

In 1962 Prien and Liske conducted a study which explored the relationship between first-level supervisor ratings, second-level supervisor ratings, and incumbent self-ratings of job performance. The ratings were carried out for research purposes and were made across eight performance items (Prien & Liske, 1962).

A small but significant average correlation of 0.25 was obtained between the ratings of the first-level supervisor and the self-ratings. The self-ratings displayed less variability and greater

leniency than did the supervisor ratings (Prien & Liske, 1962). In addition, a factor analysis of the ratings of the first-level supervisor and the incumbent resulted in a five-factor solution. The first factor was identified as a general factor, on which all performance items loaded significantly. The second factor had significant loadings on each of the supervisor ratings, and represented supervisor halo. The third factor had significant loadings on the incumbent self-ratings, and represented incumbent halo. The last two factors represented unique variance apart from rating source bias (Prien & Liske, 1962).

Kirchner (1965) similarly compared incumbent self-ratings with supervisor ratings for technical employees. Ratings were made across five performance dimensions using a 5-point graphic rating scale. The ratings were collected for research purposes only.

The results revealed greater halo in the supervisor ratings than in the incumbent self-ratings. In contrast, greater leniency was present in the self-ratings than in the supervisor ratings (Kirchner, 1965).

The construct validity of performance ratings was assessed by Lawler (1967), using the Campbell and Fiske (1959) multitrait-multimethod approach. As part of his review, Lawler (1967) compared supervisor ratings and self-ratings of management performance on three performance dimensions. Examination of the rater by dimension intercorrelation matrix revealed that comparisons between the supervisor ratings and the incumbent self-ratings resulted in

nonsignificant convergent and discriminant validity (Lawler, 1967).

Thornton (1968) believed that if self-appraisals were to be effective in eliciting an individual's cooperation, there must be agreement between the individual's self-ratings and the supervisor's ratings of that individual. Ratings were obtained from executive-level incumbents and their immediate supervisors. The ratings were made using a 5-point Likert scale to rate 27 dimensions depicting important aspects of the executive's job. The ratings were made for feedback purposes. The criterion of primary interest was an index of promotability (Thornton, 1968).

Analysis revealed little agreement between the two rating sources. The average correlation was 0.23 and was not statistically significant. Furthermore, the overall mean for self-ratings was greater than the overall mean for supervisor ratings, with those incumbents considered to be least promotable displaying the most leniency in their ratings. In contrast, the self-ratings displayed less halo error than did the supervisor ratings (Thornton, 1968).

Similar to Lawler (1967), Nealey and Owen (1970), conducted a study determining the construct validity of performance ratings of nurses using supervisors and incumbents as the two sources of ratings. The construct validity of the ratings was assessed using the Campbell and Fiske (1959) approach. Ratings were made on three dimensions of nursing performance. The results of the multitrait-multimethod analysis revealed that there was no evidence of the convergent validity or discriminant validity of the ratings (Nealey & Owen, 1970). These

results supported the findings reached by Lawler (1967).

The construct validity of supervisor ratings and self-ratings of effort and job performance for engineers was investigated by Williams and Seiler (1973). Two measures of job effort and two of job performance were used: (a) a seven-dimension work motivation (effort) scale, (b) a global measure of effort, (c) a five-dimension, behaviorally anchored rating scale of job performance, and (d) a global measure of job performance. The raters were informed that the ratings were being conducted for research purposes only (Williams & Seiler, 1973).

The results of the multitrait-multimethod analysis revealed significant convergent validity for supervisor ratings and self-ratings across both effort and performance, with greater intercorrelations for performance than for effort. Moderate levels of discriminant validity were obtained for both sets of ratings for the performance measures. In addition, greater halo was present in the supervisor ratings than in the self-ratings for both the motivation and job performance scales (Williams & Seiler, 1973).

The effects of the role of the rater on performance ratings were studied by Klimoski and London (1974). Three different sources of ratings (supervisor, peer, incumbent) were used to assess the performance of hospital nurses. The ratings were made across 19 dimensions of nursing effectiveness and one overall measure of performance. A 20-point graphic rating scale was used; the participants were informed that the ratings were for research purposes

only (Klimoski & London, 1974).

The results indicated that each rater group displayed a significant halo bias. However, the incumbent self-ratings displayed less halo and variability and more leniency than the other two groups of raters. The ratings were subjected to a hierarchical factor analysis in order to understand the underlying dimensionality of the ratings. Six factors emerged from the factor analysis. One factor was interpreted to be a general factor. Supervisor and peer ratings had high loadings on this factor, while self-ratings had low loadings. These results indicated that the difference between the rating sources was not only a difference in degree, but also a difference in the perceived dimensions evaluated by the raters (Klimoski & London, 1974). Three of the factors that emerged represented the rater sources, indicating that the three rater sources rated performance from a different perspective. The last two factors represented unique solution variance (Klimoski & London, 1974).

Heneman (1974) studied the relationship between self-ratings and supervisor ratings of managerial performance. Ratings were made on a 7-point rating scale across nine performance dimensions, including a dimension measuring overall performance. Ratings were obtained from incumbent managers and their immediate supervisors across several organizations. All ratings were used for research purposes only. Measures of leniency, variability, and halo were obtained. In addition, evidence of the construct validity of the ratings was

assessed (Heneman, 1974).

The results indicated that three of the nine self-ratings were significantly less than the corresponding supervisor ratings. Thus, unlike previous studies, these self-ratings displayed somewhat less leniency than did the supervisor ratings. The self-ratings also displayed significantly greater variability than the supervisor ratings in three instances. The supervisor ratings contained greater halo than the self-ratings. Finally, some evidence was obtained for the convergent validity and discriminant validity of the ratings (Heneman, 1974).

Baird (1977) hypothesized that the degree of congruence between self-ratings and supervisor ratings was a function of the amount of self-esteem of the incumbent and the amount of incumbent satisfaction with supervision. The participants were from various job categories, ranging from managerial to clerical positions of a state agency. The results of this study were used for research purposes only. Performance was measured using a relative rating format; each incumbent was compared to other incumbents across five performance dimensions (Baird, 1977).

The results indicated that both supervisor and incumbents displayed rating halo, but the supervisors display the greater halo. The results also revealed that the correlations between the two sources of ratings were low, indicating that the halo observed came from different points of origin (Baird, 1977). This same conclusion was reached by Klimoski and London (1974). Moreover, the group of

incumbents high on self-esteem and rated low by their supervisors, displayed the most disagreement with their supervisors. This group of incumbents rated themselves greater than did their supervisors, indicating greater rating leniency. Furthermore, high self-esteem incumbents reported greater satisfaction when they were rated high on performance by their supervisor. Those incumbents rated low reported less satisfaction. However, these results did not occur for those incumbents low on self-esteem (Baird, 1977).

The extent of leniency, halo, and differential dimensionality in peer, supervisor, and self-ratings was investigated by Holzbach (1978). Performance was measured on seven items; ratings were made using an 8-point graphic rating scale. The participants, managerial and professional employees, were informed that the data were being collected for research purposes only.

The results indicated that the self-ratings were more lenient than either the peer ratings or the supervisor ratings. In addition, significant correlations between supervisor ratings and self-ratings were obtained only for two of the performance dimensions. This is in contrast to the correlations between the supervisor ratings and the peer ratings, which were significant for each of the performance dimensions (Holzbach, 1978). A multitrait-multimethod analysis of variance was conducted to determine the construct validity of the performance ratings. Strong evidence was obtained for convergent validity, but no support was obtained for discriminant validity. The

analysis also indicated that there was significant halo present in the ratings. Finally, the underlying dimensionality of the performance ratings was determined via a principal components factor analysis. This analysis resulted in a three-factor solution. The three factors defined self-rating bias, peer rating bias, and supervisor rating bias respectively (Holzbach, 1978). This reinforces the strong rater bias found in previous studies comparing various rating sources (e.g., Klimoski & London, 1974; Prien & Liske, 1962).

Kraiger (1985) conducted a meta-analysis assessing the leniency, halo, construct validity and relative weighting of self, peer, and supervisor ratings. It was concluded that self-ratings were slightly more lenient than peer or supervisor ratings, but had less halo. In addition, low levels of convergent and discriminant validity were found between the self-ratings and the other two sources of ratings. Moreover, some evidence was obtained indicating that the different rating sources weighted the various performance dimensions differently in arriving at their evaluation of overall performance effectiveness. However, an attempt to determine the amount of variance in these ratings accounted for by the purpose of the appraisal was precluded due to too little variation (Kraiger, 1985).

Summary of Main Findings Regarding the Source of Appraisal Ratings

Supervisor ratings and incumbent self-ratings have been the focus of much attention. Research has concentrated on the psychometric qualities of the ratings provided by these two types of raters. In addition, comparisons between these two sources of ratings have been

directed at the construct validity of the obtained ratings.

Self-ratings have been found to display less halo and variability errors than supervisor ratings. In contrast, self-ratings have been found to display greater leniency error than supervisor ratings. In addition, low levels of convergent and discriminant validities have been obtained between these two sources of ratings.

The Present Study

It is clear, from the above, that a great deal of research has been devoted to understanding performance appraisal. Much attention has been directed at increasing the psychometric qualities of performance ratings. Thus, research has concentrated on reducing leniency and halo in performance ratings and increasing the variability of performance ratings. In addition, attempts have been made to assess the construct validity of performance ratings. In this regard, several factors considered to be important contributors to performance rating variance have been focused upon, such as, rater training, rating scale formats, source of ratings, and purpose of the appraisal.

Two of these factors, source of ratings and purpose of the appraisal, were of primary concern in the present study. It was proposed that the potential for the joint influence of these factors on performance ratings, although currently unexplored, was not only probable, but also theoretically justified. Thus, DeCotiis and Petit (1978) stated that the general reluctance on the part of the rater to complete the appraisal instrument accurately was due, in part, to the interaction between the purpose of the appraisal and the consequences

for the rater. Landy and Farr (1980) addressed the impact of the appraisal purpose on the rater's cognitive process, and DeNisi et al. (1984) discussed how the ratings supplied by the rater varied depending upon the purpose of the appraisal. Each provided the foundation for this researcher's proposal of an interaction between the source of ratings and purpose of the appraisal.

Furthermore, the importance of considering incumbent self-ratings and supervisor ratings of job performance, as the principle sources of ratings, was addressed in terms of clarifying job requirements and responsibilities, and increasing incumbent motivation to rate, and decreasing the defensiveness of the incumbent during the appraisal interview.

The need to examine the interaction between these two sources of ratings and the purpose of the appraisal was identified by Heneman (1974). However, a review of the literature concerning the psychometric qualities of self-ratings conducted by Thornton (1980) revealed that this interaction has remained an unexplored area. Thornton (1980) concluded that the existing data did not permit conclusions to be made as to whether the quality of self-ratings was due to the purpose of the appraisal. This same conclusion would still appear to be applicable, given that only Farh and Werbel (1986) have investigated the impact of appraisal purpose on students' self-ratings. However, these researchers investigated only one highly observable performance dimension, classroom participation, and only one dependent measure leniency. Moreover, only the dichotomy of appraisal

conducted for research purpose or administrative purpose was used.

In addition, the majority of the past research, as indicated above, was conducted for research purposes only. The results from these studies revealed that the self-ratings were more lenient and displayed less halo and variability than the corresponding supervisor ratings. In contrast, Heneman (1974), who also investigated research purposes, obtained different results. He found self-ratings to be less lenient and to display more variability than the corresponding supervisor ratings. Still further inconsistency was introduced by Farh and Werbel (1986); they found greater leniency and less variability in self-ratings when the ratings were conducted for an administrative purpose.

Therefore, given the inconsistencies of the past research examining supervisor and self-ratings and the lack of research which has systematically varied the purpose of the appraisal, the present study was conducted. This study will be the first to examine the interactive effects of appraisal purpose and source of ratings. In addition, the current study will be conducted in a field setting, using actual job incumbents (nursing assistants) and their supervisors (nurses) as participants. Another distinguishing feature of the present study is the inclusion of a control condition (no instructional set provided to the raters). With the exception of Driscoll and Goodwin (1979), no other study has included such a control condition. Since appraisal purpose has consistently been operationalized via varying instructional sets to the raters, it is important to determine

if the absence of an instructional set affects the obtained results.

The objectives of this study were to: (a) examine the impact of the purpose of the appraisal and the source of ratings on estimates of leniency, halo, and variability, and (b) determine the impact that the purpose of the appraisal has on the construct validity of supervisor and self-ratings. Specifically, incumbent (nursing assistant) self-ratings and supervisor (nurses) ratings were compared across three appraisal purposes and a control condition. The three appraisal purposes were: (a) merit pay, (b) performance improvement, and (c) research only.

Defining the Appraisal Purpose Conditions

To ensure the organizational appropriateness of the terms used and to provide a common frame-of-reference, the definitions of the appraisal purposes were developed with the help of the participating organizations. The individuals who helped in this process were not included in the actual study. These purpose manipulations were provided on a cover sheet that preceded the actual rating form.

Merit Pay. The definition for merit pay stated that based upon the results of the performance ratings the identified target ratee (nursing assistant) could possibly receive a 7% salary increase.

Performance Improvement. The definition for performance improvement stated that the results of the performance ratings would be used to determine what in-services (seminars) were needed to help increase the quality of the identified target ratee's (nursing

assistant's) job performance.

Research Only. The definition for the research only condition stated that the results of the performance ratings would be used to help develop better rating forms.

Control Condition. The control condition received no specific instructional set. Raters were simply asked to evaluate the identified target ratee's (nursing assistant's) job performance.

Research Hypotheses

Six general hypotheses were formulated. These hypotheses were based upon the above literature reviews and the objectives of this study. Predictions were made with respect to the three appraisal purposes, but none was made for the control condition due to the lack of information in past research.

1a. Self-ratings will be more lenient than supervisor ratings (Klimoski & London, 1974; Parker et al., 1959; Thornton, 1980).

1b. Self-ratings will display less halo than supervisor ratings (Klimoski & London, 1974; Parker et al., 1959; Thornton, 1980).

1c. Self-ratings will be less variable than supervisor ratings (Klimoski & London, 1974; Parker et al., 1959; Thornton, 1980).

2a. Ratings conducted for the research purpose will be less lenient than ratings conducted for either the merit pay or performance improvement purposes (Bernardin et al., 1981; Farh & Werbel, 1986; McIntyre et al., 1984). No specific hypothesis is advanced for halo.

2b. Ratings conducted for the research purpose will be more

variable than ratings conducted for either the merit pay or performance improvement purposes (Bernardin et al., 1981; Farh & Werbel, 1986; McIntyre et al., 1984). No specific hypothesis is advanced for halo.

3. Purpose of the appraisal and source of ratings will interact to affect leniency, such that relative to the supervisor ratings: (a) the greatest amount of leniency will be present in the self-ratings conducted for the merit pay purpose; (b) the least amount of leniency will be present in the self-ratings conducted for the research purpose; and (c) an intermediate amount of leniency will be present in the self-ratings conducted for the performance improvement purpose (Farh & Werbel, 1986; Thornton, 1968).

4. Purpose of the appraisal and source of ratings will interact to affect variability, such that relative to supervisor ratings: (a) the greatest amount of variability will be present in the self-ratings conducted for the research purpose; (b) the least amount of variability will be present in the self-ratings conducted for the merit pay purpose; and (c) an intermediate amount of variability will be present in the self-ratings conducted for the performance improvement purpose (Farh & Werbel, 1986; Heneman, 1974).

5. Purpose of the appraisal and source of ratings will interact to affect halo, but no specific hypotheses are advanced.

6. The construct validity of the performance ratings will be affected by the purpose of the appraisal, but no specific hypotheses are advanced.

II. METHOD

Participants

The data for the present study were collected from two nursing homes located within 10 miles of each other in Central Virginia. The participants were 168 nursing assistants and their immediate supervisors (n=43). The justification for combining the data from the two nursing homes was based upon the following: (a) both nursing facilities provided care for the same types of patients (elderly, retired, and infirm), and (b) the primary job duties and responsibilities of the nursing assistants were judged to be the same in each nursing home by the respective directors of nursing. Statistical support for the above justification was also obtained. Nursing home was coded as a study characteristic, but was not found to affect any of the dependent measures significantly, either as a main effect or as part of an interaction effect ($p > .05$).

The response rate for the nursing assistants was 90% (n=152) and the response rate for the supervisors was 91% (n=39). From these responses, 135 nursing assistant-supervisor pairs of performance ratings were formed. Sixteen nursing assistants were subsequently removed because of their failure to respond correctly to the manipulation check. The remaining 119 nursing assistant-supervisor pairs comprised the final sample in the present study. Of the 119 nursing assistants included in the final sample, 97% (n=115) were female and 3% (n=4) were male. Ninety-six percent (n=114) were white and 4% (n=5) were black. Their ages ranged from 17 to 68 with a mean

age of 33.

Of the 39 nurses included in the final sample, 97% (n=38) were female and 3% (n=1) were male. One hundred percent (n=39) were white. Their ages ranged from 23 to 66 with a mean age of 40.

Rating Scale

A 14-dimension, graphic-type rating scale was used in the present study (see Appendix B). This scale was developed by one of the nursing homes, and it was derived from the job description for the position of nursing assistant. The 14 dimensions were: (a) Quality, (b) Alertness, (c) Stability, (d) Safety Measures, (e) Job Knowledge, (f) Housekeeping, (g) Attendance, (h) Personal Appearance, (i) Restorative and Preventive Care, (j) Courtesy, (k) Initiative, (l) Cooperation and Attitude, (m) Caring and Friendliness, and (n) Overall Evaluation. Responses to the dimensions were made using a 5-level rating scale. Each level was anchored by a descriptive phrase.

Procedure

The procedure followed was the same for both nursing homes. A list of nursing assistants and their first-level supervisors (nurses) was generated. Based upon this list, dyads were formed consisting of a nursing assistant and a corresponding nurse. A necessary condition for the formation of a dyad was that the nurse had to be highly familiar with the particular nursing assistant. These formed pairs were then randomly assigned to one of four appraisal purpose conditions: (a) merit pay, (b) performance improvement, (c) research, or (d) control. The percentages of participants from the two nursing homes was in the

ratio of 40 to 60. This approximate ratio was maintained during the assignment of the dyads to the appraisal purpose conditions.

A pilot test of the study materials (instructional sets for appraisal purpose, rating form, and post-experimental questionnaire) was conducted with a group of nursing assistants from each nursing home. The pilot test was conducted to determine if the purpose manipulation was salient, and if the rating form and questionnaire were understandable. The study materials were modified according to the results of the pilot test.

Three weeks before the administration of the study materials a departmental memorandum was delivered to the nursing staff. The memorandum stated that the nursing department was conducting a joint project with the experimenter, concerning perceptions towards nursing assistant evaluations.

In the memorandum for the participants from the nursing home that developed the rating form (see Appendix C), the nursing assistants were informed that they would be asked to evaluate their own job performance using the department's current rating form. The nurses were informed that they would also be asked to evaluate nursing assistants' job performance using the same rating form as the nursing assistants. The participants were informed that the ratings provided by the nursing assistants and their nurses would be compared to determine the similarity of their perceptions of nursing assistant performance. They were told that this information would help the department determine the appropriateness of the rating form and help to meet the needs of the

nursing staff. The participants were also informed that the information they provided would be kept confidential, by the experimenter, and that the results obtained would not affect their job status.

The memorandum for the participants from the nursing home that did not develop the rating form was essentially the same (see Appendix D). The nursing assistants were informed that they would be asked to evaluate their own job performance using the provided rating form. The nurses were informed that they would also be asked to evaluate nursing assistants' job performance using the same rating form as the nursing assistants. These participants were also informed that the ratings provided by the nursing assistants and their nurses would be compared to determine the similarity of their perceptions of nursing assistant performance. The only difference was that these participants were informed that the provided rating form, although not currently used, could possibly be used, in the near future, in that department. They were then similarly told that this information would help the department determine the appropriateness of the rating form and help to meet the needs of the nursing staff. These participants were also informed that the information they provided would be kept confidential, by the experimenter, and that the results obtained would not affect their job status.

The participants (from both nursing homes) were also informed that the experimenter would be conducting group introductory sessions the following week to discuss further their role in the joint project.

The introductory sessions were held for each of the different shifts in the nursing homes. In these sessions, the experimenter discussed in more detail the nature of the project. The participants were informed that some of them would be asked to complete their evaluations for a specific administrative purpose, to be defined on the actual rating form. The specific purposes were not discussed at these sessions. The participants were told that it was important to determine how their perceptions compared for different appraisal purposes because sometimes performance ratings were used as input for more than one type of administrative decision.

The participants were then informed that the materials for the project would be distributed to them by the nursing department the following week. They were instructed to return their completed forms to the nursing department by the specified due date. The participants were given approximately 10 days to complete the project.

A sample cover sheet that preceded the actual rating form was then presented to the participants. The important information contained on the cover sheet (e.g., location of the name of the nursing assistant to be evaluated, and location of the defined appraisal purpose) was discussed. The participants were then provided with a sample item from the actual rating form. The sample item was used to illustrate how to use the rating form properly. The participants were told that the rating form would contain 14 different performance dimensions. They were informed that, like the sample item, each dimension would be followed by a definition. Under the dimension was a 5-level rating

scale. Each level described a different level of job performance for that particular dimension. They were then instructed that they were to read the respective dimension and definition and then place an "X mark" over the descriptive phrase that best represented the nursing assistant's job performance on that dimension. The participants were then reminded to rate nursing assistant job performance with respect to the appraisal purpose defined on the cover sheet to the rating form.

The nursing assistants were then informed that they would have to complete one rating form to evaluate their own job performance. The nurses were informed that because there were more nursing assistants than nurses, they would be asked to complete more than one form with each form corresponding to a different nursing assistant. The participants were then informed that they would be asked to complete a short questionnaire after they had completed their evaluations (see Appendix E). The questionnaire was designed to obtain their reactions and perceptions to performance evaluations in general and the nursing assistant rating form in particular. In addition, the questionnaire contained the purpose manipulation check.

Following this, the experimenter answered any questions the participants had concerning the administration of the project and again assured the participants that the results obtained would be kept confidential and not affect their job status. Finally, the participants were thanked for their cooperation and reminded of the return date for their evaluations. The introductory sessions lasted

approximately 20 minutes.

The following week the rating forms and questionnaires were distributed to the participants. Each rating form had a cover sheet of general instructions (see Appendixes F through M). The cover sheet, as discussed before, included information indicating the particular nursing assistant to be evaluated, the specific appraisal purpose, the definition of that purpose, and a reminder to evaluate the nursing assistant's job performance for that particular purpose.

Written instructions concerning the proper use of the rating form were included on the rating form itself. These instructions directed the rater to place an "X mark" over the scale anchor that best described the level of performance of the nursing assistant being evaluated. Similarly, written instructions concerning the proper use of the questionnaire were included on the questionnaire itself.

A follow-up postcard was sent to those participants who had not returned their completed forms by the specified return date. Those participants who had not responded to this initial follow-up were then contacted, in person, by the assistant directors of each nursing department.

Following the completion of the data analyses, the participants were debriefed. Each participant was provided with a brief overview of the purpose of the study and a summary of the major findings of the study. In addition, a formal presentation of the study findings was made to the head administrators and nursing department directors of the

host organizations.

Experimental Design

This study utilized a 4 by 2 by 13 mixed between within-subjects design (Tabachnick & Fidell, 1983). The four level, between subjects factor was appraisal purpose (i.e., merit pay, performance improvement, research, and control). The two level, within subjects factor was source of ratings (i.e., self-ratings provided by nursing assistants, and supervisor ratings provided by nurses). The thirteen level, within subjects factor was performance dimensions (overall evaluation, dimension 14, was not included in the analyses). The ratees (nursing assistants) were by design, nested within the purpose of the appraisal. Each supervisor (nurse) rated an average of 3 ratees (nursing assistants) for a given purpose.

Dependent Measures

Leniency. The definition of leniency utilized in the present study was the mean ratings across ratees within dimensions (Borman & Vallon, 1974). This was analyzed with a 4 (purpose) by 2 (source of ratings) by 13 (dimension) analysis of variance with repeated measures on the two within subjects factors.

Halo. Halo in the present study was conceptualized as the inability or failure of a rater to discriminate among the performance dimensions within a ratee (Saal, Downey, & Lahey, 1980). It was operationalized as the standard deviation across dimensions (Borman, 1975; Mount & Thompson, 1987; Warmke & Billings, 1979). Halo was analyzed with a 4 (purpose) by 2 (source of ratings) analysis of

variance with repeated measures on the source of ratings factor. In addition, a factor analysis was conducted to determine the dimensionality of the performance ratings (Holzbach, 1978; Klimoski & London, 1974).

Variability. The definition of variability utilized in the present study was the standard deviation within dimensions across ratees (Borman & Dunnette, 1975). This was analyzed with a 4 (purpose) by 2 (source of rating) by 13 (dimension) analysis of variance with repeated measures on the two within subjects factors.

Construct Validity. The definition of construct validity utilized in the present study was the degree of convergent validity (agreement between the measures in the ordering of the ratees), discriminant validity (differential ordering of the ratees by the dimensions), method bias (differential ordering of the ratees by the sources of ratings), and error (measurement and sampling error) in the performance ratings (Dickinson, 1977; 1987). This was analyzed using multitrait-multimethod analysis of variance procedures (Dickinson, 1977; 1987). In the present study, the multimethods were the sources of ratings. Thus, the analysis for construct validity may more appropriately be referred to as multitrait-multirater. A separate analysis of variance was conducted for each of the appraisal purpose conditions. This permitted comparisons of the obtained results to be made with respect to the different appraisal purposes. The psychometric interpretation of the sources of variation are summarized in Table 1.

The random effects of Ratees, Ratees x Sources, Ratees x

Table 1

Summary Table of the Psychometric Interpretations of the MTMR Design Within Each Appraisal Purpose Condition.

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

Dimensions, and Error provide the information concerning the construct validity of the ratings. Ratees depicts convergent validity, Ratees x Dimensions depicts discriminant validity, and Ratees x Sources depicts method bias (halo). Variance components and intraclass correlation coefficients (Bartko, 1966; Vaughan & Corballis, 1969) were computed for each of the sources of variation. Variance components provide a comparison of the relative sizes of convergent validity, discriminant validity, method bias, and measurement error while controlling for degrees of freedom (Dickinson, 1987). An intraclass correlation coefficient is a ratio of a source's variance component divided by the sum of all estimated variance components (Dickinson, 1987). These ratios enable comparisons of convergent validity, discriminant validity, and method bias to be made across the different appraisal purpose conditions.

Post-Experimental Questionnaire Analysis

The items on the post-experimental questionnaire were designed for three different purposes: (a) to provide information concerning the success of the appraisal purpose manipulation, (b) to provide potential explanatory information for the results of the study, and (c) to provide feedback to the host organizations. Only the responses to the items on the post-experimental questionnaire relevant to the first two purposes will be addressed.

The participants were asked to respond to two questions concerning the purpose of their appraisal ratings (i.e., questions 1 and 9 of the post-experimental questionnaire). Both questions emphasized correct

recognition of the appraisal purpose. A correct response to both of these items constituted a successful purpose manipulation. The frequencies of responses to the manipulation check are presented in Table 2. As shown in the main diagonal, 119 out of the 135 participants (88%) correctly responded to the manipulation check, $X^2(12, N = 135) = 302.44, p < .05$. Sixteen participants (12%) failed to respond correctly to the two questions.

One item asked the participants to indicate the typical use for performance information in their department (i.e., question 13 of the post-experimental questionnaire). It was believed that responses to this item would provide greater insight into the results of the current study. The frequencies of the responses to this item are presented in Table 3. As shown, 90 out of the 119 participants (76%) responded that performance information was typically used for performance improvement purposes, $X^2(9, N = 119) = 13.77, p > .05$. The nonsignificant chi-square indicates that the participants' responses to the question of the typical use of performance information were independent of their assigned to appraisal purpose condition.

The remaining items were not directly relevant to the hypotheses of the present study, but they were included to provide the host organizations with desired feedback. These items addressed issues related to the ease of use of the rating form, the comprehensiveness of the rating form, the ability to document performance using the rating form, and overall satisfaction with the rating form.

Table 2

Contingency Table of Assigned Appraisal Purpose by Perceived Appraisal Purpose.

		Assigned Purpose				
		Merit Pay	Performance Improvement	Research	Control	
Perceived Purpose	Merit Pay	36	1	0	0	37
	Performance Improvement	0	32	3	6	41
	Research	0	0	30	4	34
	Control	0	0	0	21	21
	Job Promotion	1	1	0	0	2
		37	34	33	31	

N = 135

Table 3

Contingency Table of Assigned Appraisal Purpose by Typical Use of Performance Information.

		Assigned Purpose: Merit Pay	Performance Improvement	Research	Control	
Typical Use	Merit Pay	11	1	4	1	17
	Job Promotion	0	0	0	0	0
	Performance Improvement	23	28	22	17	90
	Employee Development	2	3	4	3	12
		36	32	30	21	

N = 119

III. RESULTS

The results of this study will be presented respectively for each of the four dependent measures: (a) leniency, (b) halo, (c) variability, and (d) construct validity.

Leniency Effects. Leniency was defined as the mean ratings across ratees within dimensions. A higher mean rating indicated a greater leniency effect.

The results of the 4 x 2 x 13 ANOVA are summarized in Table 4. A significant main effect was obtained for Dimensions ($F(12, 1380) = 14.81$, $p < .01$). In addition, significant interactions were obtained for Source x Dimension ($F(12, 1380) = 16.48$, $p < .01$) and Source x Dimension x Purpose ($F(36, 1380) = 1.94$, $p < .01$). The significant main effect for Dimensions was expected. It was assumed that there would be differences among the mean ratings across the dimensions. A Newman-Keuls post hoc test was then conducted for the Dimensions effect. The results of the Newman-Keuls analysis are presented in Table 5. As shown, the means of the dimensions related to the technical aspects of the job, Quality (D1), Alertness (D2), and Job Knowledge (D5) were significantly greater than the means of the dimensions related to the interpersonal aspects of the job, Stability (D3), Courtesy (D10), Caring and Friendliness (D13), and Cooperation and Attitude (D12). Thus, greater leniency was evident in the ratings of the technical dimensions compared to the interpersonal dimensions.

The significant Source x Dimension interaction indicates that, for certain dimensions, there were differences between the self-ratings

Table 4

Analysis of Variance Summary Table for Leniency Effects.

Source	df	MS	F-Ratio
Between Subjects			
Purpose (P)	3	2.72	0.73
Ratees (R)/P	115	3.74	
Within Subjects			
Dimensions (D)	12	7.93	14.81 **
D x P	36	0.40	0.74
D x R/P	1380	0.54	
Rater Source (S)	1	0.11	0.41
S x P	3	0.02	0.06
S x R/P	115	0.28	
S x D	12	5.51	16.48 **
S x D x P	36	0.65	1.94 **
S x D x R/P	1380	0.33	

**$p < .01$.

Table 5

Newman-Keuls Post Hoc Test for Leniency Effects for the Dimensions Effect.

Dimensions												
11	3	13	10	9	12	6	4	8	5	2	7	1
—	—											
	—	—	—	—	—							
			—			—						
						—	—	—	—	—	—	—
					—			—				

Note. The dimensions are ordered by increasing mean value. The dimensions that are underscored in a row are not significantly different from each other, e.g., D11 and D3. D11 = initiative ($\underline{M}$ = 3.43); D3 = stability ($\underline{M}$ = 3.56); D13 = caring and friendliness ($\underline{M}$ = 3.66); D10 = courtesy ($\underline{M}$ = 3.74); D9 = restorative and preventive care ($\underline{M}$ = 3.74); D12 = cooperation and attitude ($\underline{M}$ = 3.77); D6 = housekeeping ($\underline{M}$ = 3.97); D4 = safety measures ($\underline{M}$ = 4.00); D8 = personal appearance ($\underline{M}$ = 4.01); D5 = job knowledge ($\underline{M}$ = 4.03); D2 = alertness ($\underline{M}$ = 4.06); D7 = attendance ($\underline{M}$ = 4.08); D1 = quality ($\underline{M}$ = 4.08).

(nursing assistants) and the supervisor ratings (nurses). This interaction was further investigated using a simple effects analysis of variance in which Rater Source was examined for each level of the Dimensions factor. The results of this analysis are summarized in Table 6. Significant differences were obtained for 9 of the 13 dimensions: (a) Quality, (b) Alertness, (c) Stability, (d) Safety Measures, (e) Job Knowledge, (f) Housekeeping, (g) Attendance, (h) Personal Appearance, and (i) Cooperation and Attitude. In each instance, the mean self-ratings were greater than the mean supervisor ratings (see Table 7). Thus, the self-ratings displayed greater leniency than did the corresponding supervisor ratings.

This finding provides partial support for Hypothesis 1a of this study. Hypothesis 1a stated that the self-ratings would be more lenient than the supervisor ratings. Complete support for this was contingent upon a significant Rater Source main effect. However, the main effect for Rater Source was not significant ($F(1, 115) = .41$). Nevertheless, the presence of the significant Source x Dimension interaction and the greater mean self-ratings do lend support to the hypothesis of greater leniency effects for the self-ratings. However, these results indicate that the relatively greater leniency of the self-ratings is not uniform across all dimensions, but dependent upon the particular set of dimensions.

There were no significant differences among the mean ratings of the appraisal purpose conditions. The Purpose effect failed to reach statistical significance ($F(3, 115) = .73$). Thus, no support was

Table 6

Simple Effects Analysis of Variance for Leniency Effects for the Rater Source x Dimension Interaction.

Source	df	MS	F-Ratio
Rater Source at D1	1	7.04	21.02 **
Rater Source at D2	1	2.83	8.44 **
Rater Source at D3	1	3.07	9.15 **
Rater Source at D4	1	2.83	8.44 **
Rater Source at D5	1	2.40	7.18 **
Rater Source at D6	1	11.36	33.92 **
Rater Source at D7	1	2.85	8.52 **
Rater Source at D8	1	1.68	5.01 *
Rater Source at D9	1	0.81	2.43
Rater Source at D10	1	1.22	3.63
Rater Source at D11	1	0.06	0.19
Rater Source at D12	1	4.02	12.01 **
Rater Source at D13	1	0.94	2.82

Note. The error term was the original error term for the Rater Source x Dimension interaction: R/P x D x S = 0.334, df = 1380. D1 = quality; D2 = alertness; D3 = stability; D4 = safety measures; D5 = job knowledge; D6 = housekeeping; D7 = attendance; D8 = personal appearance; D9 = restorative and preventive care; D10 = courtesy; D11 = initiative; D12 = cooperation and attitude; D13 = caring and friendliness.

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

Table 7

Means for the Simple Effects Analysis of Variance for Leniency Effects for the Rater Source x Dimension Interaction.

	Means	
Dimension	Nurses	Nursing Assistants
Quality	3.91	4.25
Alertness	3.95	4.12
Stability	3.45	3.67
Safety Measures	3.89	4.11
Job Knowledge	3.93	4.13
Housekeeping	3.75	4.19
Attendance	3.97	4.19
Personal Appearance	3.92	4.09
Cooperation and Attitude	3.64	3.90

obtained for Hypothesis 2 of this study. Hypothesis 2 stated that the ratings conducted for the research only purpose would be less lenient than the ratings conducted for either the merit pay or performance improvement purposes.

The significant Source x Dimension x Purpose interaction indicates that, for certain dimensions, the purpose of the appraisal did interact with the source of ratings to affect leniency. A simple effects analysis of variance was conducted to determine for which dimensions this interaction was significant. The results of this analysis are presented in Table 8. There was a significant Source x Purpose interaction for 9 of the 13 dimensions: (a) Quality, (b) Alertness, (c) Stability, (d) Safety measures, (e) Job Knowledge, (f) Housekeeping, (g) Attendance, (h) Courtesy, and (i) Cooperation and Attitude. To clarify these interaction effects, either a Newman-Keuls post hoc test or a Scheffe's multiple comparison post hoc test was performed on the Source x Purpose interaction for each of the nine dimensions. The Scheffe's test, which allows for the simultaneous testing of all contrasts, was conducted only if the Newman-Keuls test failed to uncover meaningful pair-wise differences. Four dimensions: (a) Stability, (b) Safety Measures, (c) Attendance, and (d) Courtesy required the use of Scheffe's multiple comparison test.

Quality. The self-ratings for the merit pay purpose ($\underline{M}$ = 4.31) were greater than the ratings of the supervisors for the merit pay purpose ($\underline{M}$ = 3.92). The self-ratings for the performance improvement purpose ($\underline{M}$ = 4.38) were greater than the supervisor ratings for the

Table 8

Simple Effects Analysis of Variance for Leniency Effects for the Rater Source x Purpose x Dimension Interaction.

Source	df	MS	F-Ratio
S x P at D1	7	1.31	3.93 **
S x P at D2	7	1.02	3.05 **
S x P at D3	7	0.73	2.19 *
S x P at D4	7	0.89	2.68 **
S x P at D5	7	1.08	3.24 **
S x P at D6	7	1.89	5.67 **
S x P at D7	7	2.46	7.36 **
S x P at D8	7	0.46	1.36
S x P at D9	7	0.27	0.82
S x P at D10	7	0.68	2.05 *
S x P at D11	7	0.26	0.79
S x P at D12	7	0.80	2.41 *
S x P at D13	7	0.55	1.65

Note. The error term was the original error term for the Rater Source x Dimension interaction: R/P x D x S = 0.334, df = 1380. S = rater source; P = purpose. D1 = quality; D2 = alertness; D3 = stability; D4 = safety measures; D5 = job knowledge; D6 = housekeeping; D7 = attendance; D8 = personal appearance; D9 = restorative and preventive care; D10 = courtesy; D11 = initiative; D12 = cooperation and attitude; D13 = caring and friendliness.

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

performance improvement purpose (M = 3.88).

Alertness. The self-ratings for the merit pay purpose (M = 4.14), were greater than the ratings of the supervisors for the merit pay purpose (M = 3.72).

Stability. The significant Source x Purpose interaction for this dimension was due to the linearly increasing trend present in the mean self-ratings. The respective means for the self-ratings were 3.47 for the research only purpose, 3.67 for the control condition, 3.72 for the performance improvement purpose, and 3.81 for the merit pay purpose. In contrast, relatively little variation was present in the mean ratings of the supervisors. The respective means for the supervisor ratings were 3.47 for the research only purpose, 3.48 for the control condition, 3.41 for the performance improvement purpose, and 3.44 for the merit pay purpose.

Safety Measures. The significant Source x Purpose interaction for this dimension was due to the linearly increasing trend present in the mean supervisor ratings. The respective means for the supervisor ratings were 3.67 for the research only purpose, 3.81 for the control condition, 4.00 for the performance improvement purpose, and 4.03 for the merit pay purpose. In contrast, relatively little variation was present in the mean self-ratings. The respective means for the self-ratings were 4.03 for the research only purpose, 4.05 for the control condition, 4.10 for the performance improvement purpose, and 4.22 for the merit pay purpose.

Job Knowledge. The self-ratings for the merit pay purpose (M =

4.25) were greater than the ratings of the supervisors for the merit pay purpose (M = 3.83).

Housekeeping. The self-ratings for the merit pay purpose (M = 4.28) and research only purpose (M = 4.27) were greater than the ratings of the supervisors for the merit pay purpose (M = 3.69) and research only purpose (M = 3.77).

Attendance. The significant Source x Purpose interaction for this dimension was due to the linearly increasing trend present in the mean supervisor ratings. The respective means for the supervisor ratings were 3.48 for the control condition, 3.77 for the research only purpose, 4.11 for the merit pay purpose, and 4.31 for the performance improvement purpose. In contrast, relatively little variation was present in the mean self-ratings. The respective means for the self-ratings were 4.05 for the control condition, 3.97 for the research only purpose, 4.25 for the merit pay purpose, and 4.41 for the performance improvement purpose.

Courtesy. The Scheffe's multiple comparison test failed to uncover any significant contrasts for this dimension. It would appear that although this dimension contributed to the overall significance of the Source x Purpose x Dimension interaction, it did not account for much of the total variance, which explains why no significant contrasts were found.

Cooperation and Attitude. The self-ratings for the merit pay purpose (M = 4.00) were greater than the ratings of the supervisors for

the merit pay purpose (M = 3.56).

These findings provide some support for Hypothesis 3 of this study. Hypothesis 3 stated that purpose of the appraisal would interact with source of ratings to affect leniency. In particular, relative to the supervisor ratings, the greatest amount of leniency was predicted to be present in the self-ratings for the merit pay purpose, followed by the performance improvement purpose, and then the research only purpose. Although, the Source x Purpose interaction was not significant ($F(3, 115) = .06$), the significant Source x Purpose x Dimension interaction revealed that source of ratings and purpose of the appraisal did interact. However, the significance of the interaction was dependent upon the particular set of dimensions. In addition, although the hypothesized predicted order of leniency effects was not obtained, the self-ratings for the merit pay purpose typically displayed the greatest amount of leniency.

Halo Effects. Halo was defined as the standard deviation across dimensions within ratees. A lower mean standard deviation indicated a greater halo effect.

The results of the 4 x 2 ANOVA are summarized in Table 9. As shown, a significant main effect for rater source was obtained ($F(1, 115) = 12.02$, $p < .01$). This indicates that there was a difference in the mean standard deviations between the two rater groups. In particular, the mean standard deviation for the self-ratings (M = 14.44) was greater than the mean standard deviation for the supervisor

Table 9

Analysis of Variance Summary Table for Halo Effects.

Source	df	MS	F-Ratio
Between Subjects			
Purpose (P)	3	2.86	0.78
Ratees (R)/P	115	3.69	
Within Subjects			
Rater Source (S)	1	29.79	12.02 **
S x P	3	1.10	0.45
S x R/P	115	2.48	

**$p < .01$.

ratings ($\underline{M}$ = 13.68). This finding reveals that the self-ratings exhibited less halo effect than the supervisor ratings. The job incumbents (nursing assistants) were better able than the supervisors (nurses) to differentiate among the performance dimensions. Thus, complete support was obtained for Hypothesis 1b of this study. Hypothesis 1b stated that the self-ratings would display less halo than the supervisor ratings.

In contrast, Hypothesis 5 of this study was not supported. It stated that the purpose of the appraisal would interact with the source of ratings to affect halo. The Source x Purpose interaction was not significant ($\underline{F}(3, 115) = .45$).

In addition to the ANOVA, a principal axes factor analysis with varimax rotation was conducted on the self-ratings and the supervisor ratings. This analysis provided some insight into how the two rater groups perceived the underlying relationships among the performance dimensions. The results of this analysis are presented in Table 10. Three factors emerged from the factor analysis. The first factor had significant loadings on all of the supervisor ratings. Each of the 13 dimensions had loadings of greater than .40 on this factor and near zero loadings on the other two factors. Thus, the supervisors perceived job performance to be comprised of a single factor. This factor represented supervisor halo and accounted for 25% of the variance. The next two factors had significant loadings (greater than .40) on the self-ratings and near zero loadings on the supervisor ratings. The first of these two factors (defined by the dimensions

Table 10

Principal Axes Factor Analysis (Varimax Rotation) of Supervisor (Nurse) Ratings and Self- (Nursing Assistant) Ratings.

Dimension	Factors 1	2	3	Final communality estimate
Supervisor (Nurse) Ratings				
Quality	0.780	0.180	-0.014	0.641
Alertness	0.723	0.262	-0.132	0.609
Stability	0.533	0.131	0.145	0.322
Safety Measures	0.681	0.143	-0.044	0.486
Job Knowledge	0.797	0.200	-0.118	0.689
Housekeeping	0.694	0.047	0.091	0.492
Attendance	0.473	0.162	-0.099	0.260
Personal Appearance	0.605	0.025	0.026	0.367
Restorative and Preventive Care	0.721	0.164	-0.056	0.550
Courtesy	0.750	-0.110	0.234	0.629
Initiative	0.613	0.053	-0.098	0.388
Cooperation and Attitude	0.846	-0.184	0.176	0.781
Caring and Friendliness	0.750	-0.114	0.119	0.590

Table 10 (concluded)

Dimension	Factors 1	2	3	Final communality estimate
Self- (Nursing Assistant) Ratings				
Quality	0.196	0.621	-0.001	0.424
Alertness	0.040	0.604	0.152	0.390
Stability	-0.106	0.305	0.435	0.293
Safety Measures	0.004	0.335	0.246	0.173
Job Knowledge	0.186	0.747	0.100	0.603
Housekeeping	0.064	0.276	0.109	0.092
Attendance	0.003	0.193	0.060	0.041
Personal Appearance	0.104	0.161	0.260	0.104
Restorative and Preventive Care	-0.030	0.379	0.291	0.229
Courtesy	0.144	0.062	0.647	0.443
Initiative	-0.037	0.116	0.283	0.095
Cooperation and Attitude	0.140	0.110	0.570	0.357
Caring and Friendliness	-0.035	0.108	0.717	0.527
Eigenvalue	6.467	2.134	1.972	
Percent of Variance	24.871	8.208	7.586	

Note. Factors with eigenvalues less than 1.0 were not considered. Loadings of 0.4 and above were used to define factors. Factor1 = Supervisor Halo; Factor2 = Job Task Understanding and Performance; Factor3 = Personal Qualities.

Quality, Alertness, and Job Knowledge) represented Job Task Understanding and Performance and accounted for 8% of the variance. The second of these two factors (defined by the dimensions Stability, Courtesy, Cooperation and Attitude and Caring and Friendliness) represented Personal Qualities and accounted for 7% of the variance. Thus, unlike the supervisors, the incumbents perceived job performance to be comprised of two separate factors (cf. Parker et al., 1959; Zammuto, London, & Rowland, 1982). These findings support the results of the ANOVA. The job incumbents displayed less halo, which was manifested by perceiving two distinct job performance factors. The greater supervisor halo was manifested by perceiving only one job performance factor.

Variability Effects. Variability was defined as the standard deviation within dimensions across ratees. A higher within dimension standard deviation indicated a greater variability effect.

The results of the 4 x 2 x 13 ANOVA are summarized in Table 11. As shown, a significant main effect was obtained for the Dimensions factor ($F(12, 360) = 2.35$, $p < .01$). In addition, a significant interaction was obtained for Source x Dimension ($F(12, 360) = 3.95$, $p < .01$). Once again, the significant Dimensions main effect was expected, but was not relevant to the hypotheses of this study. It was not considered in any further analysis.

The Source x Dimension interaction indicates that, for certain dimensions, there were differences in the variability of the ratings between the two rater groups. A simple effects analysis of variance

Table 11

Analysis of Variance Summary Table for Variability Effects.

Source	df	MS	F-Ratio
Between Subjects			
Purpose (P)	3	1.19	2.10
Ratees (R)/P	30	0.57	
Within Subjects			
Dimensions (D)	12	0.31	2.35 **
D x P	36	0.13	0.98
D x R/P	360	0.13	
Rater Source (S)	1	0.08	0.70
S x P	3	0.12	1.13
S x R/P	30	0.11	
S x D	12	0.46	3.95 **
S x D x P	36	0.09	0.80
S x D x R/P	360	0.12	

**$p < .01$.

was then conducted to investigate further this interaction. The results of this analysis are summarized in Table 12. Significant differences were obtained for 5 of the 13 dimensions: (a) Quality, (b) Safety Measures, (c) Initiative, (d) Cooperation and Attitude, and (e) Caring and Friendliness. For the dimensions Quality and Cooperation and Attitude, the self-ratings were less variable than the supervisor ratings. However, for the dimensions Safety Measures, Initiative, and Caring and Friendliness, the self-ratings were more variable than the supervisor ratings (see Table 13).

These findings provide mixed support for Hypothesis 1c of this study. Hypothesis 1c stated that the self-ratings would be less variable than the supervisor ratings. Complete support for this was contingent upon a significant Rater Source main effect. However, the main effect for Rater Source was not significant ($F(1, 360) = .70$). The presence of the significant Source x Dimension interaction only minimally supports this hypothesis because, even for those dimensions for which a mean variability difference was found, in less than half of these instances were the self-ratings less variable. These findings also do not support Hypothesis 4 of this study. This hypothesis stated that appraisal purpose and source of ratings would interact to affect variability. The greatest amount of variability was predicted in the self-ratings for the research only purpose, followed by the performance improvement purpose, and then the merit pay purpose. However, the Source x Purpose interaction was not significant ($F(3, 30) = 1.13$).

Construct Validity. Construct validity was evaluated using

Table 12

Simple Effects Analysis of Variance for Variability Effects for the Rater Source x Dimension Interaction.

Source	df	MS	F-Ratio
Rater Source at D1	1	1.28	11.00 **
Rater Source at D2	1	0.19	1.65
Rater Source at D3	1	0.22	1.94
Rater Source at D4	1	0.62	5.35 *
Rater Source at D5	1	0.00	0.02
Rater Source at D6	1	0.00	0.00
Rater Source at D7	1	0.07	0.62
Rater Source at D8	1	0.05	0.40
Rater Source at D9	1	0.29	2.48
Rater Source at D10	1	0.07	0.60
Rater Source at D11	1	0.78	6.71 **
Rater Source at D12	1	0.82	7.03 **
Rater Source at D13	1	0.58	4.96 *

Note. The error term was the original error term for the Rater Source x Dimension interaction: R/P x D x S = 0.116, df = 360. D1 = quality; D2 = alertness; D3 = stability; D4 = safety measures; D5 = job knowledge; D6 = housekeeping; D7 = attendance; D8 = personal appearance; D9 = restorative and preventive care; D10 = courtesy; D11 = initiative; D12 = cooperation and attitude; D13 = caring and friendliness.

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

Table 13

Means for the Simple Effects Analysis of Variance for Variability Effects for the Rater Source x Dimension Interaction.

	Means	
Dimension	Nurses	Nursing Assistants
Quality	0.69	0.42
Safety Measures	0.41	0.60
Initiative	0.59	0.81
Cooperation and Attitude	0.61	0.39
Caring and Friendliness	0.57	0.76

analysis of variance procedures (Dickinson, 1977; 1987; Kavanagh, MacKinney, & Wolins, 1971). A separate ANOVA was conducted for each of the four appraisal purpose conditions. This permitted inter-purpose comparisons of the obtained construct validity estimates. Construct validity was defined as the degree of convergent validity (Ratees effect), discriminant validity (Ratees x Dimension interaction), method bias (Ratees x Sources of Rating interaction) and Error (sampling and measurement) in the performance ratings. The results of these analyses are presented below by appraisal purpose condition.

Merit Pay. The results of the ANOVA are summarized in Table 14. There was a significant Ratees main effect ($F(35, 420) = 8.43$, $p < .01$). In addition, a significant interaction effect was obtained for Ratees x Dimension ($F(420, 420) = 1.52$, $p < .01$). These findings provide support for the convergent validity and discriminant validity of the ratings respectively. No support was obtained for the Ratees x Source interaction (i.e., method bias) ($F(35, 420) = .67$).

Performance Improvement. The results of the ANOVA are summarized in Table 15. There was a significant Ratees main effect ($F(31, 372) = 14.61$, $p < .01$). In addition, a significant interaction effect was obtained for Ratees x Dimension ($F(372, 372) = 1.61$, $p < .01$). However, the Ratees x Source interaction was not significant ($F(31, 372) = .88$). These results parallel the findings of the merit pay analysis. Evidence was obtained for convergent validity and discriminant validity, but not for method bias.

Research Only. The results of the ANOVA are summarized in Table

Table 14

Summary Table for the MTMR Analysis of Performance Ratings for the Merit Pay Purpose.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	12	3.412	6.00 **	0.039	0.059
Rater Source (S)	1	0.090	0.36	-0.000	0.000
S x D	12	2.369	6.33 **	0.055	0.083
Ratees (R)	35	3.156	8.43 **	0.107	0.161
D x R	420	0.569	1.52 **	0.098	0.148
S x R	35	0.251	0.67	-0.009	0.000
Error	420	0.374		0.374	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC = variance component; ICC = intraclass correlation coefficient.

**$p < .01$.

Table 15

Summary Table for the MTMR Analysis of Performance Ratings for the Performance Improvement Purpose.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	12	3.008	6.68 **	0.040	0.064
Rater Source (S)	1	0.100	0.41	-0.000	0.000
S x D	12	2.647	9.47 **	0.074	0.119
Ratees (R)	31	4.085	14.61 **	0.146	0.235
D x R	372	0.450	1.61 **	0.085	0.137
S x R	31	0.246	0.88	-0.003	0.000
Error	372	0.280		0.280	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC = variance component; ICC = intraclass correlation coefficient.

**$p < .01$.

16. There was a significant Ratees main effect ($F(29, 348) = 10.19$, $p < .01$). In addition, a significant interaction effect was obtained for Ratees x Dimension ($F(348, 348) = 1.79$, $p < .01$). The Ratees x Source interaction was not significant ($F(29, 348) = .88$). Again, evidence was obtained for convergent validity and discriminant validity, but not for method bias.

Control Condition. The results of the ANOVA are summarized in Table 17. These findings are, once again, similar to the results of the other analyses. Evidence was obtained for convergent validity ($F(20, 240) = 13.36$, $p < .01$), and discriminant validity ($F(240, 240) = 1.51$, $p < .01$), but not for method bias ($F(20, 240) = .98$).

Evidence for the construct validity of the ratings was obtained in each of the four appraisal purpose conditions. The ratees were differentially ordered by the dimensions (discriminant validity). This interaction is desirable. Work performance is multidimensional, and ratees are expected to differ in the amounts of the performance dimensions they demonstrate (Dickinson et al., 1986). The raters also agreed in their rank ordering of the ratees (convergent validity). The desirability of this is contingent upon the nature of the convergence. The convergence should be due to the amounts of the performance dimensions demonstrated by the ratees and not the methods or sources of the ratings (Dickinson et al., 1986). The lack of evidence for method bias (differential ordering of the ratees by the sources of ratings) in any of the appraisal purpose conditions, indicates that the rank ordering of the ratees was probably due to the dimensions and not the

Table 16

Summary Table for the MTMR Analysis of Performance Ratings for the Research Only Purpose.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	12	2.683	4.73 **	0.035	0.055
Rater Source (S)	1	0.010	0.04	-0.001	0.000
S x D	12	1.788	5.63 **	0.049	0.077
Ratees (R)	29	3.236	10.19 **	0.112	0.176
D x R	348	0.568	1.79 **	0.126	0.198
S x R	29	0.278	0.88	-0.003	0.000
Error	348	0.317		0.317	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC = variance component; ICC = intraclass correlation coefficient.

**$p < .01$.

Table 17

Summary Table for the MTMR Analysis of Performance Ratings for the Control Condition.

Source	df	MS	F-Ratio	VC	ICC
Dimensions (D)	12	0.863	1.53	0.007	0.010
Rater Source (S)	1	0.000	0.00	-0.001	0.000
S x D	12	1.137	3.05 **	0.036	0.052
Ratees (R)	20	4.983	13.36 **	0.177	0.258
D x R	240	0.563	1.51 **	0.095	0.138
S x R	20	0.367	0.98	-0.000	0.000
Error	240	0.373		0.373	

Note. If a source's variance component was negative, that value was used in the denominator to compute intraclass correlation coefficients, but the source's coefficient was set to zero. VC = variance component; ICC = intraclass correlation coefficient.

**$p < .01$.

sources of the ratings.

An index of the relative amounts of convergent validity, discriminant validity, and method bias, across the appraisal purpose conditions is provided by the intraclass correlation coefficient (ICC). The respective ICCs for each appraisal purpose are presented in Table 18. The ICCs for convergent validity ranged from highs of (.258) for the control condition and (.235) for the performance improvement condition, to a low of (.161) for the merit pay condition. Thus, the raters displayed higher levels of agreement in their rank ordering of the ratees for the performance improvement and control conditions. The ICCs for discriminant validity ranged from a high of (.198) for the research only condition to a low of (.137) for the performance improvement condition. Thus, the raters were best able to discriminate among the ratees with the dimensions for the research only condition. The lack of method bias that was reflected in the zero magnitude of the ICCs indicates that appraisal purpose did not affect the raters' ordering of the ratees. However, in each of the appraisal purpose conditions the amount of error variance was relatively high, ranging from a variance component value of .280 for the performance improvement purpose to a value of .374 for the merit pay purpose. Thus, a substantial amount of the variance in the ratings could not be attributable to either the sources of ratings or the dimensions.

Hypothesis 6 of this study stated that the purpose of the appraisal would affect the construct validity of the performance ratings. To test this hypothesis, the random effects sources of

Table 18

Intraclass Correlation Coefficients for the Random Effects Sources of Variance for the Appraisal Purpose Conditions.

	ICC			
Source	MP	PI	R	C
Convergent Validity	0.161	0.235	0.176	0.258
Discriminant Validity	0.148	0.137	0.198	0.138
Method Bias	0.000	0.000	0.000	0.000

Note. MP = merit pay; PI = performance improvement; R = research; C = control; ICC = intraclass correlation coefficient.

variance (Ratees, Ratees x Dimensions, Ratees x Sources) from each of the four ANOVAs were compared. Procedures outlined by Mosteller and Bush (cited in Rosenthal, 1984) were used to test for differences among the effects. In this procedure, the F-ratios of the random effects sources of variance were transformed to standard normal scores (Z-scores). These Z-scores were then formed into the appropriate contrasts, representing the hypothesized relationships among the effects. Finally, these contrasts were divided by their variances to form Z-tests. The Z-tests were compared to tabled values of the standard normal distribution to determine statistical significance.

In the present study, three contrasts were established: (a) the control condition effects were compared to the sum of the effects of the other three appraisal purposes, (b) the merit pay purpose effects were compared to the sum of the effects of the performance improvement and research only purposes, and (c) the performance improvement purpose effects were compared to the effects of the research only purpose. The Z-tests for these contrasts are presented in Table 19. An examination of Table 19 reveals that there were no significant differences among the effects across the appraisal purpose conditions. Thus, Hypothesis 6 of this study was not supported. Purpose of the appraisal did not affect the construct validity of the performance ratings.

Because of the disparity among the degrees of freedom across the appraisal purpose conditions, the above Z-test analysis was also conducted using a balanced sample design. This was accomplished by randomly deleting cases across the appraisal purpose conditions.

Table 19

Z-Tests of the Formed Contrasts of the Across Purpose Random Effects Sources of Variance for Construct Validity.

Contrast	CV	DV	MB
1) Control condition versus sum of other conditions	-1.189	-1.378	0.713
2) Merit pay condition versus sum of performance improvement and research only conditions	-1.081	-0.447	-0.881
3) Performance improvement condition versus research only condition	1.881	-0.543	0.016

Note. CV = convergent validity; DV = discriminant validity; MB = method bias.

The results of this analysis similarly revealed no significant differences among the effects.

Comparison with other MTMR Studies

To provide a context for the construct validity results of the present study, the ICC values obtained were compared to previous MTMR studies. Table 20 presents a comparison between the mean ICC values from five other studies and the present study. These five studies were selected because they permitted the comparison between self-ratings and supervisor ratings that were not confounded by other sources of ratings. The ratings were conducted for non-administrative purposes. The paucity of research that has investigated construct validity by appraisal purpose precludes the ability to compare the present across purpose findings with similar past research. The ICC values presented were either obtained from Dickinson et al. (1986) or computed using the mean squares reported in the study's summary table. In either instance, the ICC values were computed according to Bartko's (1966) definition (i.e., the ratio of a source's variance component to the sum of all relevant variance components).

The mean discriminant validity obtained in the present study ($\underline{M}$ = .155) was greater than the mean discriminant validity in the other studies ($\underline{M}$ = .088). Comparable convergent validities were obtained in the present study ($\underline{M}$ = .208) and the other studies ($\underline{M}$ = .243). Moreover, while relatively high amounts of method bias were present in the other studies ($\underline{M}$ = .282), no method bias was obtained in the present study ($\underline{M}$ = .000). In sum, these findings indicate that higher

Table 20

Comparisons of ICC Values Derived from Previous MTMR Studies.

Study	Convergent Validity	Discriminant Validity	Method Bias
Baird (1977)[a]	0.352	0.026	0.515
Heneman (1974)[a]	0.202	0.098	0.190
Mount (1984)	0.111	0.205	0.157
Prien & Liske (1962)	0.269	0.086	0.241
Steel & Ovalle (1984)	0.279	0.029	0.306
Mean ICC Values Across Studies	0.243	0.088	0.282
Mean ICC Values For Present Study	0.208	0.155	0.000

[a] Intraclass correlation coefficient (ICC) obtained from Dickinson et al. (1986).

discriminant validity and lower method bias were obtained in the present study compared to previous studies. Nevertheless, these collective ICCs may be considered to be of low to moderate magnitude (see Dickinson et al., 1986).

IV. DISCUSSION

Models of performance appraisal have identified several factors which influence performance ratings. The factors of source of ratings and purpose of the appraisal were of primary interest in the present study. It was proposed that these two factors would interact to affect the psychometric qualities (leniency, halo, variability) and construct validity of performance ratings. Two sources of ratings (self-ratings and supervisor ratings) and four appraisal purposes (merit pay, performance improvement, research only, and a control condition) were included in the present study.

Six general research hypotheses were generated concerning the influence of the two study variables on the psychometric properties and construct validity of performance ratings. This discussion addresses the viability of these hypotheses, provides explanations for the obtained results, and where appropriate, integrates the results with past research findings. A separate discussion will be presented for each of the dependent variables, followed by an overall conclusion.

Leniency Effects. Leniency was operationalized by comparing the mean ratings of different rating sources. A greater mean rating indicated a leniency effect. The majority of past research has demonstrated that self-ratings are typically greater than corresponding supervisor ratings (e.g., Holzbach, 1978; Kirchner, 1965; Klimoski & London, 1974; Kraiger, 1985; Prien & Liske, 1962; Thornton, 1968). Past research has also demonstrated that ratings conducted for administrative purposes are typically more lenient than ratings

conducted for non-administrative purposes (e.g., Aleamoni & Hexner, 1980; Bernardin et al., 1981; Borresen, 1967; Farh & Werbel, 1986). In the present study it was hypothesized that: (a) self-ratings would be more lenient than supervisor ratings, (b) ratings conducted for merit pay or performance improvement purposes would be more lenient than ratings conducted for research only purposes, and (c) the greatest amount of leniency would be present in the self-ratings for merit pay purposes and the least amount would be present in the self-ratings for research only purposes. An intermediate amount of leniency was predicted to be present in the self-ratings for performance improvement purposes.

The hypothesis (1a) that the self-ratings would be more lenient than the supervisor ratings was partially supported. Although the Rater Source main effect was not significant, the Source x Dimension interaction was significant. The self-ratings were more lenient for 9 of the 13 performance dimensions. This finding indicates that the greater tendency of self-ratings to be more lenient than supervisor ratings is not necessarily a uniform phenomenon. Whether the self-ratings will be more lenient depends on the particular set of performance dimensions. This dimension dependence is consistent with previous research (e.g., Holzbach, 1978; Mount, 1984; Thornton, 1968). In each of these studies a significant Source x Dimension interaction was obtained. Although this interaction is commonly found, the lack of overlap in the dimensions included in one study to another makes explanation of this interaction somewhat difficult (Kraiger, 1985).

One framework which may aid in understanding the differential effect of performance dimensions on the leniency of self-ratings is presented by Festinger (1954). In his social comparison theory it is proposed that two different motivational forces are present in self-evaluations. One motivational force directs the individual to obtain accurate self-evaluation information. The other motivational force directs the individual to obtain inflated self-evaluation information. The question then becomes, what factors determine which of these motivational forces will be dominant? One such factor may be the performance dimension set itself. Different dimensions may evoke one or the other of these motivational forces, leading to greater leniency effects for some dimensions and not others.

In the present study, the factor analysis of the self-ratings and supervisor ratings resulted in two factors emerging for the self-ratings (see Table 10). One factor (Job Task Understanding and Performance) was related to the technical aspects of the job and the other (Personal Qualities) was related to the interpersonal aspects of the job. A Newman-Keuls post hoc test of the Dimensions effect (see Table 5) revealed that the means of the dimensions related to the technical aspects of the job were significantly greater than the means of the dimensions related to the interpersonal aspects of the job. This lends some support to the proposition that dimension content differentially affects the quality of self-ratings. In particular, technically oriented dimensions may motivate the rater to obtain inflated self-evaluation information, while interpersonally oriented

dimensions may motivate the rater to obtain accurate self-evaluation information (Festinger, 1954).

Clearly, more research is needed addressing this issue. Research needs to identify (a) which specific types of dimension content influence performance rating, and (b) if this influence is similar across different rater sources. Is the dichotomy of technical and interpersonal content sufficient to explain the variance in the quality of performance ratings? Do dimensions that are more objective result in less rating errors? Do dimensions perceived to be more closely tied to a reward structure result in greater inflation of self-ratings? In addition, tests of the viability of Festinger's (1954) social comparison theory as a means for explaining the interactive effects of dimension content and rating source are warranted. Do performance dimensions affect the quality of self-ratings through their ability to stimulate one or the other of the two sets of motivational forces present in self-evaluations?

It was also predicted that the ratings conducted for merit pay and performance improvement purposes would be more lenient than ratings conducted for research only purposes (Hypothesis 2a). Prior research (e.g., Bernardin et al., 1981) indicated that performance ratings carried out for administrative purposes were more lenient than ratings carried out for non-administrative purposes. It is believed that this phenomenon was due to the increased consequences that the ratings for administrative purposes have for both the rater and ratee (DeCotiis & Petit, 1978). Consequently, the greatest amount of leniency was

predicted to occur for the appraisal purpose expected to hold the greatest consequences for the raters and/or ratees, the merit pay purpose. A performance improvement purpose was expected to be of less consequence for the raters and/or ratees, and so, less leniency was predicted to occur in this condition. Finally, the least consequential appraisal purpose was expected to be the research only purpose. Thus, this conditions was predicted to be the least lenient.

The results obtained did not support this hypothesis. There were no significant mean differences among the appraisal purpose conditions. Of particular interest, however, was the fact that there was no significant difference between the control condition and the other appraisal purpose conditions. Although no hypothesis was made with respect to the control condition, previous research did reveal that compared to explicit appraisal purposes, a control condition (no defined appraisal purpose) received the lowest mean ratings (Driscoll & Goodwin, 1979). One possible reason why this did not occur in the present study may be that the raters in the control condition supplied their own appraisal purpose for making their ratings. When the raters were asked to indicate the purpose for which job performance information was typically used in their department, approximately 76% of the participants responded, performance improvement (see Table 3). It would seem plausible to assume that in the absence of any defined purpose that the raters made their ratings for purposes for which they were most familiar, in this case, performance improvement. This would explain the nonsignificant difference between the control condition and

the other appraisal purpose conditions.

Research has demonstrated that under non-administrative conditions, self-ratings are more lenient than supervisor ratings (Holzbach, 1978; Klimoski & London, 1974). However, Farh and Werbel (1986) found that when both administrative and non-administrative conditions are present, self-ratings display greater leniency under the administrative condition. It was thus predicted that appraisal purpose and source of ratings would interact to affect leniency (Hypothesis 3).

Some support was obtained for this hypothesis. The results revealed a significant Rater Source x Purpose x Dimension interaction. A significant Rater Source x Purpose interaction occurred for 8 of the 13 performance dimensions. In all but three of these instances the differences between the sources of ratings were due to the greater mean self-ratings for the merit pay purpose. For one dimension (i.e., Housekeeping), the self-ratings for research only purposes significantly exceed the supervisor ratings. Thus, similar to Farh and Werbel (1986), the greater leniency effect for self-ratings was observed predominantly under the administrative conditions and not the research only condition. Replication of these findings is needed.

Future research attempts may benefit by systematically varying the administrative conditions under which the ratings occur according to a criterion such as their perceived importance to the raters. Obtaining this kind of information a priori, and then selecting administrative conditions from different levels of this continuum, may provide greater

insight into the role that appraisal purpose plays in performance ratings. Furthermore, the moderating role played by the dimensions factor again suggests that future research needs to address the effect of dimension content on the interaction between rating source and appraisal purpose.

Halo Effects. Halo was defined as the failure of a rater to discriminate among performance dimensions. It was operationalized as the standard deviation across dimensions. The smaller the standard deviation the greater the halo effect. Past research has consistently demonstrated that self-ratings display less halo than supervisor ratings (e.g., Heneman, 1974; Parker et al., 1959; Prien & Liske, 1962). This finding has typically been attributed to the different perspective each rating source has of the target position. It was predicted that the self-ratings would be subject to less halo effect than the supervisor ratings (Hypothesis 1c). This hypothesis was supported. A significant Rater Source main effect was obtained for the halo measure. The self-ratings were more variable across the performance dimensions than were the supervisor ratings.

To determine if an alternate job perspective explanation was appropriate, a factor analysis of the performance ratings was conducted. If different factors emerged for the sources of ratings, an alternate job perspective explanation would be supported. The results clearly indicated that the two sources of ratings did not perceive the target position similarly. The supervisors perceived job performance to be comprised of a single factor. In contrast, the nursing

assistants perceived job performance to be comprised of two distinct factors, one related to the technical aspects of the job and the other to the interpersonal aspects of the job. This may be explained by the greater intimacy the nursing assistants have regarding their own jobs compared to the nurses. Although, both groups work closely together, the nurses are still removed from all the daily routines of the nursing assistants. This distance may preclude the nurses from recognizing the subtleties of the nursing assistant position. The nurses would have more limited information to base their evaluations on than would the nursing assistants. An availability bias (Tversky & Kahneman, 1974) would be operating, the nurses would be rating performance based upon those activities most familiar to them. Consequently, the nurses would be inclined to rate job performance from a more stereotypic, global perspective than would the nursing assistants.

The hypothesis that source of ratings and purpose of appraisal would interact to affect halo was not supported. The Rater Source x Purpose interaction was not significant. Regardless of the purpose for making performance ratings, self-ratings displayed more discrimination among performance dimensions than supervisor ratings. This finding and the significant Rater Source main effect suggests that job incumbents may be in the best position to judge their own strengths and weaknesses (Thornton, 1980). Thus, organizations might benefit by including self-ratings for purposes such as of employee development, determination of training needs, and career development.

Variability Effects. Variability was defined as the extent to

which the ratings discriminate among the ratees within dimensions. It was operationalized as the standard deviation across ratees within dimensions. It is expected that not all ratees would perform at the same level on a particular dimension, and this should be manifested as a relatively large within dimension standard deviation. However, past research has typically demonstrated that self-ratings display less within dimension variability than supervisor ratings (e.g., Klimoski & London, 1974; Parker et al., 1959; Prien & Liske, 1962). It would appear that although job incumbents are better able to discriminate among their own levels of performance across dimensions (halo), they are less able than supervisors to discriminate among each other's level of performance within dimensions (variability).

Self-ratings were predicted to be less variable than supervisor ratings (Hypothesis 1c). Mixed support was obtained for this hypothesis. Although, the rater source main effect was not significant, the Rater Source x Dimension interaction was significant. Variability differences between the sources of ratings were found at 5 of the 13 dimensions. However, in three of these instances (Safety Measures, Initiative, and Caring and Friendliness) the self-ratings were more variable than the supervisor ratings. For these three dimensions the self-ratings were better able than the supervisor ratings to discriminate among the performance levels of the nursing assistants.

Research that has included dependent measures of both leniency and variability has revealed that these two measures tend to covary

negatively; conditions of greatest leniency also display least variability (Farh & Werbel, 1986; Klimoski & London, 1974; Parker et al., 1959; Prien & Liske, 1962). Consequently, it was also predicted that more variability should be present in the ratings for research purposes than for either merit pay or performance improvement purposes (Hypothesis 2b). Additionally, it was predicted that source of ratings and purpose of appraisal should interact to affect variability (Hypothesis 4).

No support was obtained for Hypothesis 2b. There were no differences in variability across the different appraisal purpose and control conditions. Similarly, no support was obtained for Hypothesis 4. Source of ratings and appraisal purpose did not interact to affect variability. Nevertheless, these findings do partially support the frequently observed negative relationship between leniency and variability. Of the three dimensions that displayed greater self-rating variability, two of these dimensions did not display greater self-rating leniency. In contrast, the two dimensions that did display less self-rating variability, also displayed greater self-rating leniency. Thus, consistent with other research, greater variability was associated with less leniency. This implies that leniency effects may be reduced by either directly decreasing the rater's motivation to be lenient or by indirectly increasing the variability of the rater's responses (Farh & Werbel, 1986).

Construct Validity. Construct validity was defined as the degree of convergent validity, discriminant validity, and method bias present

in the performance ratings. These terms were operationalized according to the analysis of variance procedures of Kavanagh et al. (1971) and Dickinson (1977; 1987). In general, research comparing self-ratings with supervisor ratings has revealed evidence of moderate levels of convergent validity, low levels of discriminant validity, and high levels of method bias (e.g., Heneman, 1974; Steel & Ovalle, 1984). However, very few attempts have been made to assess the construct validity of performance ratings for different appraisal purposes. In the vast majority of studies, performance ratings were collected only for non-administrative purposes. The influence that appraisal purpose may have on construct validity is not known (see Dickinson et al., 1986; Kraiger, 1985). Therefore, this study represented a first attempt at systematically varying appraisal purpose to determine its effect on construct validity.

Although, the three contrasts tested did not reveal any significant differences in the construct validity estimates of the appraisal purpose conditions, the results obtained do present some interesting insights.

Higher levels of convergent validity were obtained in the performance improvement purpose and the control condition than in the merit pay and research purposes. This contrast was tested a posteriori but was not found to be significant. Nevertheless, this observation does warrant discussion. The higher levels of convergent validity in the performance improvement purpose and control condition may have occurred because they were less cognitively demanding than the merit

pay and research purposes. Baker (1986) observed higher levels of convergent validity for an assigned role leaderless group discussion compared to a non-assigned role leaderless group discussion. This was partially attributed to the greater cognitive demands that the non-assigned role placed on the raters. All else being equal, unfamiliar tasks tend to be more cognitively demanding for individuals than familiar tasks. As stated before, the participants indicated that performance information was typically collected for performance improvement purposes. It is believed that this accounted for the highly similar ICCs obtained for the performance improvement purpose and control condition (see Table 18). The participants in the control condition were assumed to be rating performance for that purpose most familiar to them, performance improvement. The greater familiarity that the participants had with performance improvement purposes implies that the rating tasks for the performance improvement purpose and control condition were less cognitively demanding than the rating tasks for either merit pay or research purposes, resulting in the higher levels of convergent validity in these two conditions. Future research addressing the impact of cognitive demand on convergent validity is needed.

In terms of discriminant validity, lower levels were obtained for the merit pay and performance improvement purposes than for the research purpose (again the ICC for the control condition was highly similar to that for the performance improvement purpose). This may have resulted because of the greater perceived consequences associated

with these two administrative purposes. Research suggests that ratings conducted for research purposes are more accurate than ratings conducted for administrative purposes (McIntyre et al., 1984; Murphy et al., 1984). This may be due to the greater ability and/or motivation of raters to discriminate among the amounts of dimensions demonstrated by the ratees for research purposes (discriminant validity). To avoid any negative consequences associated with ratings for administrative purposes, the raters may avoid differentially ordering the ratees on the dimensions. By doing so, the raters may believe that there is a reduced probability that their ratings will be challenged and that they will be required to justify the ratings.

No method bias (i.e., differential ordering of the ratees by the sources of ratings) was obtained in the present study. The Ratees x Rater Source effect was not significant for any of the appraisal purpose conditions. This is in contrast to past research that has typically found evidence of moderate to high levels of method bias. The presence of this effect is believed to be due to differential opportunities to observe performance (Dickinson et al., 1986). This effect may not have been observed in the present study because of the relatively high amount of contact between the nursing assistants and the nurses. Nursing assistants are responsible for reporting to their nursing supervisors on a daily basis. In addition, often in the care of the patients, the nursing assistants and nurses perform their respective duties concurrently, increasing the opportunity for the nurse to observe nursing assistant performance. Future research would

benefit by examining the effects of frequency and relevancy of job contact and prior job experience on rater source method bias.

Conclusions

Research comparing different rater sources has, for the most part, concentrated on the comparison between supervisor ratings and peer ratings (Mount, 1984). Much less emphasis has been placed upon the study of self-ratings as an alternative rating source. When self-ratings have been compared to supervisor ratings, the self-ratings have been determined to be more lenient, less variable, and subject to less halo (Thornton, 1980). Estimates of the construct validity of the ratings of these two sources have revealed moderate levels of convergent validity, low levels of discriminant validity, and high levels of method bias (e.g., Prien & Liske, 1962; Steel & Ovalle, 1984). However, the majority of this research has been carried out for strictly non-administrative purposes. Very little research has been conducted examining the effects of appraisal purpose on the quality of performance ratings (Dickinson et al., 1986; Harris & Schaubroeck, 1988). The present study was conducted to determine the effects of appraisal purpose and source of ratings on the psychometric properties of performance ratings (leniency, halo, variability), and the construct validity of performance ratings (convergent validity, discriminant validity, and method bias).

In general, the results of this study revealed that compared to supervisor ratings, self-ratings were more lenient and subject to less halo. Mixed findings were obtained for variability estimates; self-

ratings were less variable for some dimensions, but more variable for other dimensions. Appraisal purpose did affect performance ratings in terms of leniency. The significant Rater Source x Purpose X Dimension interaction indicated that the higher self-ratings carried out for either merit pay or performance improvement purposes accounted for much of the mean differences between the self-ratings and the supervisor ratings.

Contrasts of the obtained estimates of construct validity for the different appraisal purposes did not reveal any significant differences. Nevertheless, higher levels of convergent validity were obtained in the performance improvement purpose and control condition compared to the merit pay and research purposes. Lower levels of discriminant validity were obtained in the merit pay purpose, performance improvement purpose, and control condition compared to the research purpose. The former results were attributed to the reduced cognitive demand that the performance improvement purpose and control condition placed on the raters. The latter results were attributed to the negative consequences associated with ratings for administrative purposes. The lack of method bias in the different appraisal purpose conditions was attributed to the relatively high level of job contact between nurses and nursing assistants.

As with any research study, certain design compromises were evident that affected the obtained results. An unavoidable contaminant in the present study design was the explicit research context surrounding the ratings. The raters and ratees were aware that the

ratings would not actually affect their job status. The raters were asked to role-play and assume that the ratings they supplied would be used for the respective appraisal purpose. The consequences of the ratings were not real. McIntyre et al. (1984) proposed that appraisal purpose had its greatest affect on the emotionality of the rater, causing the rater to look beyond the short term consequences of the ratings. The rater who believes that the ratings will be used for administrative purposes recognizes the life consequences of the ratings. In the present study, the potential of the purpose manipulation to evoke the necessary emotional reactions in the raters was contingent upon the ability and motivation of the rater to successfully role-play. The inevitable variability across the participants to successfully role-play most probably accounted for the failure to observe any mean rating differences across the appraisal purpose conditions.

Continued research examining the effects of appraisal purpose on the quality of performance ratings is clearly needed. The impact that appraisal purpose has on the accuracy and validity of performance ratings is one such area, identified by Dickinson (1987). For instance, do ratings conducted for administrative purposes result in reduced levels of discriminant validity (Dickinson et al., 1986)? The findings of the present study indicated that the ratings for administrative purposes did not result in reduced estimates of discriminant validity. More research is needed examining the relationship between appraisal purpose and construct validity.

Future research must maximize the perceived consequences of performance ratings associated with a particular appraisal purpose for both the rater and ratee. The impact that the ratings could have on the job status of the ratees need to be explicitly communicated to the raters. In addition, the potential consequences that the ratings could have on the life situation of the ratees (e.g., ratee self-esteem, family life, etc.) need to be communicated to the raters. These types of contextual factors would help the defined appraisal purpose to evoke the necessary emotional reactions in the raters.

In addition, the results of the present study need to be replicated using different populations of participants. There was a high degree of contact between the nursing assistants and the nurses in the present study. This is an atypical situation. Whether these same results would or would not be obtained using job incumbents and supervisors that did not have such a high degree of job contact is unknown and needs to be addressed.

Finally, the role of dimension content on the quality of performance ratings is another area needing more empirical investigation. Does specific dimension content differentially motivate job incumbents to seek out accurate or inflated performance information? Do certain types of dimensions interact with certain appraisal purposes to affect the quality of performance ratings?

V. REFERENCES

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instruction on student course and instructor evaluation. Instructional Science, 9, 67-84.

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

Baker, T. A. (1986). Multitrait-multimethod analysis of performance ratings using behaviorally anchored and behavioral checklist formats. Unpublished masters thesis, Old Dominion University, Norfolk, VA.

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

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

Berkshire, J. R., & Highland, R. W. (1953). Forced-choice performance rating: A methodological study. Personnel Psychology, 6, 355-378.

Bernardin, H. J., & Abbott, J. (1985). Predicting (and preventing) differences between self and supervisory appraisals. Personnel Administrator, 30, 151-157.

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

Bernardin, H. J., Orban, J. A., & Carlyle, J. J. (1981). Performance rating as a function of trust in appraisal and rater individual differences. Proceedings of the Academy of Management, 311-315.

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. (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. (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., & 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.

Borresen, H. A. (1967). The effects of instructions and item content on three types of ratings. Educational and Psychological Measurement, 27, 855-862.

Burnaska, R. F., & Hollmann, T. D. (1974). An empirical comparison of the relative effects of rater response bias 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, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

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.

Cascio, W. F. (1982). Applied psychology in personnel management (2nd ed.). Reston, VA: Reston.

Centra, J. A. (1976). The influence of different directions on student ratings of instruction. Journal of Educational Measurement, 13, 277-282.

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

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

Dickinson, T. L. (1977). The analysis of variance model for multi-trait-multimethod investigations. Unpublished manuscript, Colorado State University.

Dickinson, T. L. (1987). Designs for evaluating the validity and accuracy of performance ratings. Organizational Behavior and Human Decision Processes, 40, 1-21.

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 Base Human Resources Laboratory.

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65, 147-154.

Driscoll, L. A., & Goodwin, W. L. (1979). The effects of varying information about use and disposition of results on university students, evaluations of faculty and courses. American Educational Research Journal, 16, 25-37.

Farh, J. L., & Werbel, J. D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. Journal of Applied Psychology, 71, 527-529.

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

French, W. L. (1982). The personnel management process (5th ed.). Boston, MA: Houghton Mifflin.

Gmelch, W. H., & Glasman, N. S. (1977). The effect of purposes on student evaluation of college instructors. Educational Research Quarterly, 2, 45-55.

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

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.

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

Hobson, C. J., Mendel, R. M., & Gibson, F. W. (1981). Clarifying performance appraisal criteria. Organizational Behavior and Human Performance, 28, 164-181.

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

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

Kay, E., Meyer, H. H., & French, J. R. P. (1965). Effects of threat in a performance appraisal interview. Journal of Applied Psychology, 49, 311-317.

Kirchner, W. K. (1965). Relationships between supervisory and subordinate ratings for technical personnel. Journal of Industrial Psychology, 3, 57-60.

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

Kraiger, K. (1985). Analysis of relationships among self, peer, and supervisory ratings of performance (F49620-85-C-0013). Brooks AFB, TX: Manpower and Personnel Division, Brooks Air Force Base Human Resources Laboratory.

Landy, F. J. (1985). Psychology of work behavior (3rd ed.). Homewood, IL: The Dorsey Press.

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

Landy, F. J., & Farr, J. L. (1983). The measurement of work performance: Methods, theory, and applications. New York: Academic Press.

Landy, F. J., & Trumbo, D. A. (1980). Psychology of work behavior (rev. ed.). Homewood, IL: The Dorsey Press.

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.

Locher, A. H., & Teel, K. S. (1977). Performance appraisal: A survey of current practices. Personnel Psychology, 56, 245-247; 254.

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.

Meier, R. S., & Feldhusen, J. F. (1979). Another look at Dr. Fox: Effect of stated purpose for evaluation, lecturer expressiveness, and density of lecture content on student ratings. Journal of Educational Psychology, 71, 339-345.

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

Mount, M. K., & Thompson, D. E. (1987). Cognitive categorization and quality of performance ratings. Journal of Applied Psychology, 72, 240-246.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. Journal of Educational Psychology, 76, 45-54.

Murphy, K. R., Herr, B. M., Lockhart, M. C., & Maguire, E. (1986). Evaluating the performance of paper people. Journal of Applied Psychology, 71, 654-661.

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

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.

Prien, E. P., & Liske, R. E. (1962). Assessments of higher level personnel: III. Rating criteria: A comparative analysis of supervisory ratings and incumbent self-ratings of job performance. Personnel Psychology, 15, 187-194.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills: Sage.

Rothe, H. F. (1978). Output rates among industrial employees. Journal of Applied Psychology, 63, 40-46.

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 scale: An evaluation. Organizational Behavior and Human Performance, 18, 19-35.

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

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

Spool, M. D. (1978). Training programs for observers of behavior: A review. Personnel Psychology, 31, 853-888.

Steel, R. P., & Ovalle, N. K. (1984). Self-appraisal based upon supervisory feedback. Personnel Psychology, 37, 667-685.

Tabachnick, B. G., & Fidell, L. S. (1983). Using multivariate statistics. New York: Harper & Row.

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-appraisal of job performance. Personnel Psychology, 33, 363-371.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

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

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

Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. Organizational Behavior and Human Decision Processes, 35, 314-339.

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.

Woods, S. B. (1987). The influence of rater training, scale format, and rating justification on the quality of performance ratings by three rater sources. Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

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

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

VI. APPENDIX A:

Summary of Studies Included in the Literature Reviews of Purpose of the Appraisal and Source of Appraisal Ratings

Table A-1

Characteristics of Studies Examining the Interactive Role of the Purpose of the Appraisal.

Study	Warmke & Billings (1979)	Zedeck & Cascio (1982)
Independent Variables		
Purpose	Experimental & administrative.	Development, merit pay, or retention.
Source of Ratings	Head nurses, & assistant head nurses.	Undergraduates.
Other Factors	Rater training.	Rater training.
Dependent Variables		
Halo	Greater halo in administrative purpose. Experimental purpose: Scale construction training least halo error. Administrative purpose: No training effect.	------
Leniency	No leniency effect.	------
Variability	Experimental purpose: Scale construction, & lecture training groups most variable ratings. Administrative purpose: No training effect.	Merit pay purpose: least variable ratings.
Construct Validity	------	------

Table A-1 (continued)

Study	McIntyre, Smith, & Hassett (1984)	Bernardin, Orban, & Carlyle (1981)
Independent Variables		
Purpose	Research, hiring, or feedback.	Feedback or promotion.
Source of Ratings	Undergraduates.	Police sergeants.
Other Factors	Rater training.	Rater trust & cognitive complexity.
Dependent Variables		
Halo	Too little halo across all purposes. Frame-of-reference training closest to true halo.	------
Leniency	Research purpose: least lenient.	Promotion purpose: most lenient. High Trust condition: least lenient.
Variability	------	------
Construct Validity	------	------

Table A-1 (concluded)

Study	Williams, DeNisi, Blencoe, & Cafferty (1985)	Farh & Werbel (1986)
Independent Variables		
Purpose	Salary increase, promotion, or training referral.	Research or course grade.
Source of Ratings	Undergraduates.	Undergraduates.
Other Factors	Relative or absolute rating decision.	Expectation of ratings validation.
Dependent Variables		
Halo	------	------
Leniency	Training purpose: most lenient.	Course grade purpose: most lenient.
	Absolute rating decision: most lenient.	Low expectation of validation: most lenient.
Variability	------	Course grade & low expectation of validation condition: least variable ratings.
Construct Validity	------	------

Table A-2

Characteristics of Studies Examining the Relationship Between Supervisor and Self-Ratings.

Study	Parker, Taylor, Barrett, & Martens (1959)	Prien & Liske (1962)
Independent Variables		
Purpose	Research.	Research.
Source of Ratings	Supervisors & incumbents of clerical positions.	Supervisors & incumbents.
Dependent Variables		
Halo	Self-ratings less halo.	Both rater groups displayed halo.
Leniency	Self-ratings more lenient.	Self-ratings more lenient.
Variability	Self-ratings less variable.	Self-ratings less variable.
Construct Validity	------	------

Table A-2 (continued)

Study	Kirchner (1965)	Lawler (1967)
Independent Variables		
Purpose	Research.	Research.
Source of Ratings	Supervisors & incumbents of technical positions.	Supervisors & incumbents of management positions.
Dependent Variables		
Halo	Self-ratings less halo.	------
Leniency	Self-ratings more lenient.	------
Variability	------	------
Construct Validity	------	Little evidence of either convergent or discriminant validity.

Table A-2 (continued)

Study	Thornton (1968)	Nealey & Owen (1970)
Independent Variables		
Purpose	Feedback.	Research.
Source of Ratings	Supervisors & incumbents of executive positions.	Supervisors & incumbents of nursing positions.
Dependent Variables		
Halo	Self-ratings less halo.	-------
Leniency	Self-ratings more lenient.	-------
Variability	------	------
Construct Validity	------	Little evidence of either convergent or discriminant validity.

Table A-2 (continued)

Study	Williams & Seiler (1973)	Klimoski & London (1974)
Independent Variables		
Purpose	Feedback.	Research.
Source of Ratings	Supervisors & incumbents of engineering positions.	Supervisors, peers, & incumbents of nursing positions.
Dependent Variables		
Halo	Self-ratings less halo.	Self-ratings less halo than other sources.
Leniency	------	Self-ratings more lenient than other sources.
Variability	------	Self-ratings less variable than other sources.
Construct Validity	High convergent validity across both measures of effort and performance; moderate discriminant validity for performance measure.	------

Table A-2 (continued)

Study	Heneman (1974)	Baird (1977)
Independent Variables		
Purpose	Research.	Research.
Source of Ratings	Supervisors & incumbents of management positions.	Supervisors & incumbents of positions ranging from managerial to clerical.
Dependent Variables		
Halo	Self-ratings less halo.	Self-ratings less halo.
Leniency	Self-ratings less lenient.	-------
Variability	Self-ratings more variable.	-------
Construct Validity	Some convergent and discriminant validity.	------

Table A-2 (concluded)

Study	Holzbach (1978)	Kraiger (1985)
Independent Variables		
Purpose	Research.	Literature review: meta-analysis.
Source of Ratings	Supervisors, peers, & incumbents of managerial-professional positions.	Supervisors, peers, & incumbents.
Dependent Variables		
Halo	Halo displayed for all sources.	Self-ratings less halo than other sources.
Leniency	Self-ratings more lenient.	Self-ratings more lenient.
Variability	------	------
Construct Validity	High convergent validity, no discriminant validity.	Little convergent or discriminant validity.

VII. APPENDIX B:

Rating Form: Nursing Assistant Position

RATING FORM: NURSING ASSISTANT POSITION

INSTRUCTIONS:

Listed below are a number of traits and characteristics that are important and necessary to function as a nursing assistant. PLACE AN " X " MARK ON EACH RATING SCALE, OVER THE DESCRIPTIVE PHRASE WHICH MOST NEARLY DESCRIBES THE PERSON YOU ARE RATING.

QUALITY: is the completeness of duties performed that can be relied upon with inspection - bathing, feeding, dressing, mouth care, toilet care, Temp., Pulse, Respiration, Blood Pressure and intake and output.

work frequently incomplete, must be redone	careless work- rushes just to get done	usually complete, needs only average number of reminders; a careful worker	requires little supervision; is complete and precise most of the time	requires absolute minimum of supervision; is almost always complete

ALERTNESS: is the ability to grasp instructions and directions, to observe changing conditions for signs and symptoms and reports to Charge Nurse.

slow to "catch on;" not observant	requires more than average instruction and explanation; sometimes observes signs and symptoms and reports to nurse	grasps instruction with average ability and observes patient	usually quick to "catch on;" usually observes signs and symptoms and reports to nurse	exceptionally keen and alert; very observant of signs and symptoms and is sure to report to nurse

STABILITY: is the ability to withstand pressure and to remain calm in crisis situations.

goes to "pieces" under pressure; is "jumpy" and nervous	occasionally "blows up" under pressure; is easily irritated	has average tolerance for crises; usually remains calm	copes with most pressure; very good tolerance for crises	thrives under pressure; really enjoys solving problems and crises

ATTENDANCE: is faithfulness in coming to work daily and conforming to work hours; is not tardy. Accepts schedules and follows regulations in notification of absences, vacations and sickness.

often absent without good excuse and/or frequently is tardy; does not follow regulations for notification	lax in attendance and for reporting for work on time; sometimes does not give proper notification	usually present and on time; usually gives proper notification	very prompt; regular in attendance; gives proper notification	always regular and prompt; conscientious about proper notification and volunteers for extra commitment

PERSONAL APPEARANCE: is the personal impression an individual makes on others. (Consider cleanliness, body odor, grooming, neatness and appropriateness of dress on the job in relation to the dress code.)

very untidy; ignores dress code	sometimes untidy and careless about personal appearance	generally neat and clean; satisfactory personal appearance; usually follows dress code	careful about personal appearance clean and neat; follows dress code	extremely neat and clean; always adheres to dress code

RESTORATIVE AND PREVENTIVE CARE: is to use comfort measures for the patients' positioning, exercises, walking, fluids, etc. and to record accurately and in appropriate places.

seldom bothers with preventative and and comfort measures	requires frequent reminders to do comfort measures and must be reminded to do exercises and walk patients and to record such	usually uses comfort measures as required and will walk and exercise patients and records care given	careful to use comfort measures and to do restorative care; records accurately and appropriately	extremely conscientious to do restorative and preventative care; records precisely

COURTESY: is the polite attention an individual gives to patients, staff and others.

blunt; discourteous; nasty at times	sometimes tactless	agreeable and pleasant	always very polite and willing to help	inspiring to others in being courteous and very pleasant

SAFETY MEASURES: is to provide a safe environment for the patient by using learned safety procedures; proper use of side rails, restraints, Hoyer Lift, proper body mechanics for lifting.

has little regard for safety measures and procedures	lax in using safety measures; forgets to raise side rails and use and check restraints; does not bother with Hoyer Lift	usually practices safety procedures and checks restraints every 30 minutes and releases every 2 hours	careful to use proper safety measures and procedures; is alert to safety hazards; uses and checks on restraints properly	extremely careful to use safety measures and procedures and is observant of safety hazards

JOB KNOWLEDGE: is the knowledge and understanding of duties, functions, procedures, treatments and terminology necessary for satisfactory job performance.

does not have a working knowledge of the job; poorly informed	job knowledge is limited to the the simplest duties; continually needs instruction	has sufficient knowledge to perform the job satisfactorily	has a more than adequate knowledge of the job; understands phrases and carries through	has complete and thorough knowledge of the job and can totally be relied upon

HOUSEKEEPING: is the orderliness and cleanliness in which the work gets done, condition of the patient rooms and cleaning duties. Maintains aseptic conditions during performance of duties.

disorderly or untidy; patients appear neglected; rooms messy, neglects cleaning duties	some tendency to be careless and untidy; needs to be reminded to do cleaning duties	usually keeps work rooms and patients rooms fairly neat; patients fairly neat	quite conscientious about neatness and and cleanliness; patients rooms neat	unusually neat, clean and orderly; patients very neat as well as rooms

INITIATIVE: is expression of having new ideas, for finding new and better ways of doing things and being imaginative. Contributes suggestions and ideas for patient care. Exhibits commitment for personal growth.

rarely shares a new idea; not imaginative; rarely attends meetings	occasionally has a new idea and attends meetings	has average imagination and sometimes contributes ideas; usually attends meetings	frequently suggests new ways of doing things; is very imaginative and frequently attends meetings	continually seeks new and better ways of doing things; is extremely imaginative and continues to contribute ideas; attends all meetings

COOPERATION AND ATTITUDE: is the individual's ability to work with others. Manner in which employee reacts to supervision, co-workers, patients and employee's personal suitability for the job.

works poorly with others; frequently grumbles about policy, work assignment, etc. personality unsatisfactory for this job; cannot accept criticism	attitude needs improving, is occasionally uncooperative; personality questionable for this job; difficult to accept criticism	works well with others; usually has a good attitude; person satisfactory for his job; can accept criticism	works very well with others; willing to assist others in their work; very desirable personality for this job; wants to know ways to improve performance	goes out of the way to be cooperative; excellent attitude; outstanding personality for this job; welcomes constructive criticism

CARING AND FRIENDLINESS: is the sociability, warmth and concern employee imparts in their behavior towards patients, co-workers and those who supervise.

aloof and distant; little regard for others	approachable; takes a while to warm up to others	warm, friendly and sociable; spends time with patients	very sociable and out-going; tries to involve patients in care and activities	extremely sociable; very warm and deeply caring

OVERALL EVALUATION:

very poor	poor, but improving	average	above average	outstanding

VIII. APPENDIX C:

Introductory Memorandum: Bridgewater Home

Date: January 22, 1988

To: Nursing Staff

From: Pearl Parks, R.N., Director of Nursing

Subject: Your participation and cooperation in a project concerned with the evaluation of the performance rating form for nursing assistants.

The department of nursing in cooperation with Mr. Rick Tannenbaum from the Personnel Research Laboratory of Old Dominion University is conducting a project looking at your perceptions towards performance evaluations in general and in particular about the rating form for nursing assistants used here at Bridgewater Home.

The project involves looking at how nursing assistants view their own job performance and comparing that to how nurses view nursing assistants' job performance when both groups use the same rating form. The information we obtain by looking at these shared perceptions will greatly help us to determine if the rating form for the position of nursing assistant is helping us meet the needs of our staff.

Please be assured that in no way will your responses be used against you or negatively affect you or your job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

On Wednesday, February 3rd, Mr. Tannenbaum will meet with you in groups during your normal shifts to discuss the project in more detail and to answer any of your questions.

On Wednesday, February 10th, you will receive a packet containing rating forms and an opinion questionnaire. Please promptly complete all the forms and return them back to the nursing department no later than Monday, February, 22.

Since you have the greatest understanding of the position of nursing assistant it is very important that everyone who receives a packet complete the forms. Each one of your opinions, and comments is vital to making this effort a successful one, helping us to meet your needs.

Thank you for your participation and valuable time!!!

Sincerely,

Pearl Parks, R.N.

IX. APPENDIX D:

Introductory Memorandum: Oak Lea

Date: January 22, 1988

To: Nursing Staff

From: Kim Fridinger, R.N., Assistant Director of Nursing

Subject: Your participation and cooperation in a project concerned with the evaluation of a performance rating form for nursing assistants.

The department of nursing in cooperation with Mr. Rick Tannenbaum from the Personnel Research Laboratory of Old Dominion University is conducting a project looking at your perceptions towards performance evaluations in general and in particular about a rating form for nursing assistants which may be used here at Oak Lea.

The project involves looking at how nursing assistants view their own job performance and comparing that to how nurses view nursing assistants' job performance when both groups use the same rating form. The information we obtain by looking at these shared perceptions will greatly help us to determine if the rating form for the position of nursing assistant is suitable for helping us meet the needs of our staff.

Please be assured that in no way will your responses be used against you or negatively affect you or your job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

On Thursday, February 4th, Mr. Tannenbaum will meet with you in groups during your normal shifts to discuss the project in more detail and to answer any of your questions.

On Wednesday, February 10th, you will receive a packet containing rating forms and an opinion questionnaire. Please promptly complete all the forms and return them back to the nursing department no later than Monday, February, 22.

Since you have the greatest understanding of the position of nursing assistant it is very important that everyone who receives a packet complete the forms. Each one of your opinions, and comments is vital to making this effort a successful one, helping us to meet your needs.

Thank you for your participation and valuable time!!!

Sincerely,

Kim Fridinger, R.N.

X. APPENDIX E:

Post-Experimental Questionnaire

OPINION QUESTIONNAIRE

INSTRUCTIONS:

Please complete this questionnaire AFTER you have finished making your ratings. Respond to these items based upon your reactions, views and opinions to the rating form you used and to performance evaluations in general.

1. What was the PURPOSE of your ratings AS STATED ON YOUR WRITTEN INSTRUCTIONS? (CIRCLE YOUR ANSWER)

Merit Pay	Research of Forms	Performance Improvement	Job Promotion	Purpose not Stated

2. To what extent do you believe that you would change the ratings you made if you were asked to supply ratings for a different purpose? (CIRCLE A NUMBER)

1.....2.....3.....4.....5

very little — to some extent — very great

3. To what extent do the items on the rating form reflect the important aspects of the job of nursing assistant? (CIRCLE A NUMBER)

1.....2.....3.....4.....5

very little — to some extent — very great

4a. ANSWER ONLY IF YOU ARE A NURSING ASSISTANT. To what extent do you believe that the ratings you supplied will agree with the ratings supplied by your supervisor? (CIRCLE A NUMBER)

1.....2.....3.....4.....5

very little — to some extent — very great

4b. ANSWER ONLY IF YOU ARE A NURSE. To what extent do you believe that the ratings you supplied will agree with the ratings supplied by your nursing assistant(s)? (CIRCLE A NUMBER)

1.....2.....3.....4.....5

very little — to some extent — very great

5. To what extent do you believe that it was easy to use the rating form to rate job performance? (CIRCLE A NUMBER)

1.....2.....3.....4.....5

very little — to some extent — very great

6. Who do you believe should supply job performance ratings for purposes of employee development? (CHECK MARK YOUR CHOICE)

   The employee himself/herself _____
   The employee's supervisor _____
   Both the employee and the supervisor _____

7. How much previous experience have you had as a rater of an individual's job performance? (CIRCLE A NUMBER)

   1.....2.....3.....4.....5
   very little some a great deal

8. Who do you believe should supply job performance ratings for purposes of merit pay and/or promotions? (CHECK MARK YOUR CHOICE)

   The employee himself/herself _____
   The employee's supervisor _____
   Both the employee and the supervisor _____

9. Which of the choices below is the same as the PURPOSE of your ratings? (CHECK MARK YOUR CHOICE)

   Merit Pay _____
   Performance Improvement _____
   Research of Forms _____
   Job Promotion _____
   None of the above _____

10. To what extent does the rating form allow you to adequately document an individual's job performance? (CIRCLE A NUMBER)

    1.....2.....3.....4.....5
    very little to some extent very great

11. Please RANK ORDER the below, from 1 to 4, in terms of how important a decision affecting each is to you.
    (1=LEAST important to 4=MOST important)

    Merit Pay _____ Promotion _____

    Employee Development _____ Performance Improvement _____

12. Overall, how satisfied are you with the rating form you used to rate job performance?

    1.....2.....3.....4.....5
    not at all somewhat completely

13. For what purpose is job performance information typically used in this department? (CIRCLE YOUR CHOICE)

Merit Pay ..... Promotions ..... Performance Improvement ..... Employee Development

14. Please list below what you consider to be the GOOD POINTS and the BAD POINTS of the rating form you used.

BAD POINTS	:	GOOD POINTS
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	
	:	

(THANK YOU FOR COMPLETING THIS QUESTIONNAIRE!!!)

XI. APPENDIX F:

Cover Sheet To Rating Form: Self-Ratings For Merit Pay Purpose

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORM and QUESTIONNAIRE. First Complete the rating form and then after that the questionnaire. Both of these materials need to be completed.

*** Please EVALUATE YOUR OWN JOB PERFORMANCE.

****

* The PURPOSE OF YOUR RATINGS is for A MERIT PAY INCREASE. *
*
* A MERIT PAY INCREASE means that based upon your performance *
* ratings you could possibly receive a 7% salary increase. *
*
* PLEASE RATE PERFORMANCE WITH A "MERIT PAY INCREASE" PURPOSE IN *
* MIND. *
*

****

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to look at your ratings in the future if you wish to.

PLEASE NOTE that in no way will these ratings actually affect your job or any outcomes affecting your job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

XII. APPENDIX G:

Cover Sheet To Rating Form: Self-Ratings For Performance Improvement Purpose

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORM and QUESTIONNAIRE. First complete the rating form and then after that the questionnaire. Both of these materials need to be completed.

*** Please EVALUATE YOUR OWN JOB PERFORMANCE.

****

* The PURPOSE OF YOUR RATINGS is for PERFORMANCE IMPROVEMENT. *
*
* PERFORMANCE IMPROVEMENT means that your ratings will be used to *
* determine what in-services are needed to help increase the *
* quality of your job performance. *
*
* PLEASE RATE PERFORMANCE WITH A "PERFORMANCE IMPROVEMENT" PURPOSE *
* IN MIND. *
*

****

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to look at your ratings in the future if you wish to.

PLEASE NOTE that in no way will these ratings actually affect your job or any outcomes affecting your job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENTY BY FEBRUARY 22.

XIII. APPENDIX H:

Cover Sheet To Rating Form: Self-Ratings For Research Only Purpose

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORM and QUESTIONNAIRE. First complete the rating form and then after that the questionnaire. Both of these materials need to be completed.

*** Please EVALUATE YOUR OWN JOB PERFORMANCE.

****
* The PURPOSE OF YOUR RATINGS is for RESEARCH OF RATING FORMS. *
*
* RESEARCH OF RATING FORMS means that your ratings will be used *
* to develop better rating forms. *
*
* PLEASE RATE PERFORMANCE WITH A "RESEARCH OF RATING FORMS" *
* PURPOSE IN MIND. *
*
****

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to look at your ratings in the future if you wish to.

PLEASE NOTE that in no way will these ratings actually affect your job or any outcomes affecting your job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

XIV. APPENDIX I:

Cover Sheet To Rating Form: Self-Ratings For Control Condition

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORMS and QUESTIONNAIRE. First complete each rating form and then after that the questionnaire. All of these materials need to be completed.

*** PLEASE EVALUATE YOUR OWN JOB PERFORMANCE.

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to refer back to these ratings in the future.

PLEASE NOTE that in no way will these ratings actually affect your job or the nursing assistant's job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

XV. APPENDIX J:

Cover Sheet To Rating Form: Supervisor Ratings For Merit Pay Purpose

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORMS and QUESTIONNAIRE. First complete each rating form and then after that the questionnaire. All of these materials need to be completed.

PLEASE EVALUATE THE JOB PERFORMANCE OF: _____

****
* The PURPOSE OF YOUR RATINGS is for A MERIT PAY INCREASE. *
*
* A MERIT PAY INCREASE means that based upon the performance *
* ratings you supply the above nursing assistant could possibly *
* receive a 7% salary increase. *
*
* PLEASE RATE PERFORMANCE WITH A "MERIT PAY INCREASE" PURPOSE IN *
* MIND. *
*
****

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to refer back to these ratings in the future.

PLEASE NOTE that in no way will these ratings actually affect your job or the nursing assistant's job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

XVI. APPENDIX K:

Cover Sheet To Rating Form: Supervisor Ratings For Performance Improvement Purpose

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORMS and QUESTIONNAIRE. First complete each rating form and then after that the questionnaire. All of these materials need to be completed.

PLEASE EVALUATE THE JOB PERFORMANCE OF: _____

****
* The PURPOSE OF YOUR RATINGS is for PERFORMANCE IMPROVEMENT. *
*
* PERFORMANCE IMPROVEMENT means that the ratings you supply will *
* used to determine what in-services are needed to help increase *
* the quality of the above nursing assistant's job performance. *
*
* PLEASE RATE PERFORMANCE WITH A "PERFORMANCE IMPROVEMENT" *
* PURPOSE IN MIND. *
*
****

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to refer back to these ratings in the future.

PLEASE NOTE that in no way will these ratings actually affect your job or the nursing assistant's job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

XVII. APPENDIX L:

Cover Sheet To Rating Form: Supervisor Ratings For Research Only Purpose

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORMS and QUESTIONNAIRE. First complete each rating form and then after that the questionnaire. All of these materials need to be completed.

PLEASE EVALUATE THE JOB PERFORMANCE OF: _____

****
* The PURPOSE OF YOUR RATINGS is for RESEARCH OF RATING FORMS. *
*
* RESEARCH OF RATING FORMS means that the ratings you supply will *
* be used to help develop better rating forms. *
*
* PLEASE RATE PERFORMANCE WITH A "RESEARCH OF RATING FORMS" *
* PURPOSE IN MIND. *
*
****

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to refer back to these ratings in the future.

PLEASE NOTE that in no way will these ratings actually affect your job or the nursing assistant's job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

XVIII. APPENDIX M:

Cover Sheet To Rating Form: Supervisor Ratings For Control Condition

Your Name: _____ Date: _____

This packet contains the materials you need for the rating form evaluation project outlined in the memorandum you received dated January 22. This project was further explained by Mr. Tannenbaum during your shift-time.

PLEASE READ ALL OF THESE INSTRUCTIONS VERY CAREFULLY

This packet contains your RATING FORMS and QUESTIONNAIRE. First complete each rating form and then after that the questionnaire. All of these materials need to be completed.

PLEASE EVALUATE THE JOB PERFORMANCE OF: _____

Since it is important to hear from everyone, we would like you to place your name at the top of this page. This will help to keep track of the flow of paper, and also allow you to refer back to these ratings in the future.

PLEASE NOTE that in no way will these ratings actually affect your job or the nursing assistant's job. We are only interested in your opinions and views so that we can continue to meet the needs of our staff.

Your participation in this effort will make the information we receive the best it can be. Thank you for taking the time to complete the forms!!!!!

*** RETURN YOUR COMPLETED FORMS AND QUESTIONNAIRE TO THE NURSING DEPARTMENT BY FEBRUARY 22.

ABSTRACT

PERFORMANCE APPRAISAL RATINGS AS A FUNCTION OF SOURCE

OF RATINGS AND PURPOSE OF THE APPRAISAL

Richard J. Tannenbaum
Old Dominion University, 1988
Director: Dr. Terry L. Dickinson

This study investigated the effects of purpose of appraisal ratings and source of appraisal ratings on four dependent measures: (a) leniency, (b) halo, (c) variability, and (d) construct validity. The purpose factor was comprised of four different levels: (a) merit pay, (b) performance improvement, (c) research only, and (d) no defined appraisal purpose. The rating source factor was comprised of two different levels: (a) incumbent self-ratings, and (b) supervisor ratings. One hundred and nineteen nursing assistants provided the self-ratings, and 39 nurses provided the supervisor ratings. Both sets of ratings were made using an in-house developed, 13-dimension graphic-type rating scale. Analysis of variance procedures were used to test the effects of appraisal purpose and rating source on the dependent measures. Significant Rater Source x Dimension, and Rater Source x Purpose x Dimension effects were obtained for the leniency measure. These findings provided partial support for the hypotheses that the self-ratings would be more lenient than the supervisor ratings and that rater source would interact with appraisal purpose to affect leniency. No support was obtained for the hypothesized main effect of appraisal purpose on leniency. A significant Rater Source effect was obtained for the halo measure. This finding provided complete support for the

hypothesis that the self-ratings would display less halo than the supervisor ratings. In addition, a factor analysis of the performance ratings resulted in a three-factor solution. One factor had significant loadings on all of the supervisor ratings and it represented supervisor halo. The next two factors had significant loadings on the self-ratings. The first of these factors represented Job Task Understanding and Performance. The second factor represented Personal Qualities. No support was obtained for the hypothesized interaction of appraisal purpose and rater source on halo. A significant Rater Source x Dimension effect was obtained for the variability measure. This finding provided mixed support for the hypothesis that the self-ratings would be less variable than the supervisor ratings. The self-ratings were less variable than the supervisor ratings for less than half of the dimensions examined. No support was obtained for the hypothesized interaction of appraisal purpose and rater source on variability. Significant convergent validity (Ratees effect), and discriminant validity (Ratees x Dimension interaction) were obtained for each of the appraisal purpose conditions. No method bias (Ratees x Rater Source interaction) was obtained for any of the appraisal purpose conditions. However, the hypothesis that appraisal purpose would differentially affect construct validity was not supported. The significance of these findings and recommendations for future research examining the role of appraisal purpose were discussed.