

AN ABSTRACT OF THE THESIS OF

Jamie C. Kieffer for the Master of Science

in Psychology presented on July 7, 1997

Title: An Examination of the Interrater Agreement Between Self- and
Supervisory Performance Ratings in a Subjective Occupation

Abstract approved: Brian Schrader

Employees need to have feedback about their work performance through a formal performance appraisal system. A performance appraisal can be defined as the process of evaluating employees on multiple job-related dimensions. Most organizations utilize some type of formal performance appraisal to evaluate an employee's performance on the job. Traditionally, these performance evaluations have consisted of supervisors rating their subordinates on multiple work-related dimensions. However, several studies have indicated some inherent problems with this type of evaluation. Therefore, organizations are increasingly utilizing a combined ratings method of obtaining multiple raters, including self-ratings, to improve their performance appraisal system. The present study examined the effects of six differential comparison standards (ambiguous, internal, absolute, relative-inside, relative-outside, and multiple) on the level of agreement between self- and supervisory performance ratings within the context of a subjective occupation. Forty-five self-supervisor dyads evaluated three work performance dimensions using the comparison standards. Results supported the effects of these differential comparison standards on

significantly increase when raters were using similar comparison standards.

Various supported hypotheses and research implications are discussed.

**AN EXAMINATION OF THE INTERRATER AGREEMENT
BETWEEN SELF- AND SUPERVISORY PERFORMANCE
RATINGS IN A SUBJECTIVE OCCUPATION**

Thesis

Presented to

the Division of Psychology and Special Education

EMPORIA STATE UNIVERSITY

In Partial Fulfillment

of the Requirements of the Degree

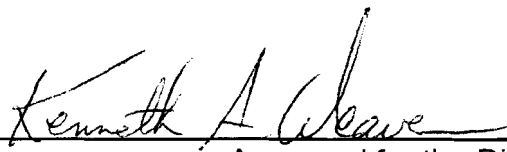
Master of Science

by

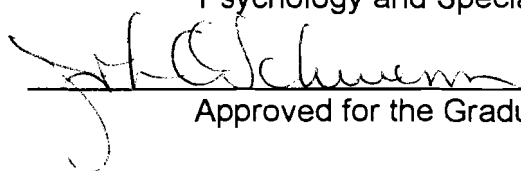
Jamie Kieffer

August 1997

1/15/17
1/17/17
1/17/17



Approved for the Division of
Psychology and Special Education



Approved for the Graduate Council

ACKNOWLEDGMENTS

As I reflect upon my graduate studies, I would like to express my deepest gratitude to several individuals. First, I would like to extend my sincere appreciation to my parents, Jim and Karen Martin, and my brother, Chris Martin, who have always been there for me. They will never know how grateful I am for our remarkable relationships. I could not have become who I am without their unconditional love, support, and friendship. I would also like to express my love and appreciation to my husband, Patrick, for his friendship, understanding, and unflinching humor. His endless emotional encouragement and considerable patience has enabled me to complete my graduate studies.

My sincere gratitude also goes to my thesis chair, Dr. Brian Schrader, for all his patient guidance throughout this process. I am thankful for his significant contributions in the development of this research and for providing me with continuous support and assistance. I also want to thank the members of my thesis committee, Dr. Stephen Davis and Dr. Janice Hoshino, for their invaluable expertise and helpful comments which added to the quality of this paper.

Finally, I must thank the rest of my family members and all of my friends who provided me with a steady source of emotional support and considerate advice during the exhausting demands of graduate school. Because there are too many of them to mention here, they will remain anonymous in this acknowledgment, but not in my heart. Thank you.

TABLE OF CONTENTS

| | |
|---------------------------------|-----------|
| ACKNOWLEDGMENTS | iii |
| TABLE OF CONTENTS | iv |
| LIST OF TABLES | vi |
| LIST OF FIGURES | vii |
| <u>CHAPTER</u> | |
| 1 INTRODUCTION | 1 |
| Review of the Literature | 3 |
| Hypotheses | 19 |
| 2 METHOD | 23 |
| Participants | 23 |
| Procedure | 25 |
| Measures | 25 |
| 3 RESULTS | 31 |
| Hypotheses | 31 |
| Hypothesis 1 | 31 |
| Hypothesis 2 | 37 |
| Hypothesis 3 | 39 |
| Hypothesis 4 | 39 |
| 4 DISCUSSION | 44 |
| Interpretation of Results | 45 |
| Limitations | 51 |

| | |
|--|----|
| Implications and Future Research | 52 |
| REFERENCES | 55 |
| APPENDICES | 61 |
| Appendix A: Approval Letter From Institutional Review Board | 61 |
| Appendix B: Transmittal Letter | 63 |
| Appendix C: Packet Instructions and Informed Consent Document | 66 |
| Appendix D: Rating Instructions and Performance Dimensions | 69 |
| Appendix E: Faculty Member Rating Sheets | 71 |
| Appendix F: Chairperson Rating Sheets | 78 |
| Appendix G: Post-Rating Comparison Standard Questions | 82 |
| Appendix H: Demographics and Comprehension Question | 86 |

LIST OF TABLES

| <u>Table</u> | | <u>Page</u> |
|--------------|--|-------------|
| 1 | Repeated Measures Analysis of Variance for Performance Ratings: Rater Source x Performance Dimension x Comparison Standard | 32 |
| 2 | Tukey's HSD Analysis of Differences Between Comparison Standard Means For Full Sample..... | 36 |
| 3 | Pearson Correlations Between Faculty Member Ratings and Chairperson Ratings for Each Comparison Standard..... | 38 |
| 4 | Observed and Expected Frequencies for Previously Used Comparison Standards Reported by Chairpersons..... | 40 |
| 5 | Means and Standard Deviations for Faculty Member-Chairperson Preference Ratings | 42 |

LIST OF FIGURES

| <u>Figure</u> | | <u>Page</u> |
|---------------|--|-------------|
| 1 | Comparison Standard x Performance Dimension Interaction | 34 |

CHAPTER 1

INTRODUCTION

Most organizations utilize some type of formal performance appraisal to evaluate an employee's performance on the job (Locher & Teel, 1988).

Traditionally, these performance evaluations have consisted of supervisors rating their subordinates on multiple work-related dimensions. Several studies, however, have indicated some inherent problems with this type of evaluation. Organizations are therefore increasingly utilizing a combined ratings method of obtaining multiple raters, including self-raters, to improve their performance appraisal system.

Although there are many advantages to using self-evaluations, previous research suggests the interrater agreement between self- and supervisory ratings is low to moderate (Landy & Farr, 1980; Thornton, 1980). Although there are inconsistent results, several studies have cited leniency error and halo error as possible explanations for this lack of convergence (Mabe & West, 1982; Meyer, 1980). Previous research also examined the different referents individuals use in performance appraisals. Goodman (1974) concluded that when individuals are evaluating their earnings, they compare themselves to "others." Specifically, they compare themselves to co-workers within their organization (relative-inside) or to individuals with similar jobs who work outside their institution (relative-outside). In a more recent study, Schrader and Steiner

(1996) concluded that raters select different comparison standards when evaluating an employee's work performance. These researchers defined a comparison standard as a referent choice that "represents the benchmark, or standard, against which a rater compares the ratee's performance" (p. 814). They modified Goodman's (1974) typology of referents and used four different comparison standards (internal, absolute, relative, and multiple) in their study. Their sample included employees who use objective, quantifiable measures for determining work success. The results of Schrader and Steiner's study indicated that when both supervisory and self-raters were provided with explicit instructions for which comparison standard to use, the interrater agreement between self- and supervisory ratings significantly increased.

The utility of subjective performance evaluations is becoming a common practice among organizations (Guba & Lincoln, 1989). This type of appraisal is different than objective evaluations as subjective performance appraisals require individuals to use judgmental, qualitative criteria. Therefore, raters must rely on their own values of what constitutes successful performance on the job (Hoffman, Nathan, & Holden, 1991). In this type of subjective rating process, it is likely different raters will use different comparison standards.

An issue which has not been delineated in the performance appraisal literature is the use of differential comparison standards in performance appraisals for subjectively based occupations. Therefore, one of the purposes

of the present study was to extend the number of comparison standards used in Schrader and Steiner's (1996) study to include internal, absolute, relative-inside, relative-outside, and multiple referents. In addition, this study utilized a completely different type of sample whose participants are traditionally evaluated on subjective performance dimensions. The results of this study contributed to the understanding of the interrater disagreement between self- and supervisory ratings in subjective occupation types.

Review of the Literature

Employees need to have feedback about their performance on the job, and formal performance appraisal systems provide this information. A performance appraisal can be defined as the process of evaluating employees on multiple work-related dimensions (Ilgen, Barnes-Farrell, & McKellin, 1993). Although most organizations have always had some form of performance evaluation, these procedures have usually been informal (Berry & Houston, 1993). However, over the years, companies have turned towards utilizing formal performance evaluation systems. In a study of 324 organizations, 94% reported using formal procedures to evaluate their employee's performance (Locher & Teel, 1988).

Purposes of a Performance Appraisal

Many companies are implementing formal performance appraisals because the evaluations can serve several purposes. A survey of 106

industrial/organizational psychologists employed in a variety of organizations suggested 20 purposes of a performance appraisal. A factor analysis of their results indicated four general uses of performance appraisals: (a) to identify strengths and weaknesses of employees, (b) to provide performance feedback, (c) to administer salaries, and (d) to document personnel decisions (Cleveland, Murphy, & Williams, 1989). All of these purposes can benefit the employees as well as the employers.

Although companies have been adopting these formal processes, they have also been searching for ways to improve their system for evaluating employee performance. The success or failure of a performance appraisal system depends on how it is implemented and how it is perceived by the participants. Traditionally, performance evaluations consisted of supervisors rating their employees on work behavior (Murphy & Cleveland, 1991), however, research cited some inherent problems in using only one source. Although it could be argued that supervisors are most familiar with employee's performance, research has shown supervisory ratings are susceptible to biases (Cascio, 1991; Landy & Farr, 1980; Thornton, 1980). Some of the biases noted in the literature include supervisors giving higher ratings to same-sex subordinates, ineffective supervisors providing less reliable and valid ratings, interaction-oriented supervisors producing more lenient ratings, and the job experience of the supervisor positively affecting the quality of ratings (Landy, 1989). Bassett and

Meyer (1968) have also suggested that supervisory ratings are substantially more time-consuming and costly for organizations. Lastly, supervisors may have a limited opportunity to observe a subordinate's work performance, and they may not produce accurate ratings of their employee's performance (Cascio, 1991). These weaknesses of supervisory ratings have led many organizations to apply an alternative approach to performance evaluations utilizing multiple raters, including self-raters.

Self-ratings Research

To improve their performance evaluations, organizations are increasingly adopting the use of the combined ratings method of obtaining multiple raters, including self-raters (Farh, Werbel, & Bedeian, 1988). Research has indicated both advantages and disadvantages to the use of self-ratings in performance evaluations, and some studies have evaluated the validity of self-ratings. Although the results of prior studies are inconclusive, there is a considerable amount of correlational research on the role of self ratings.

Advantages of self-ratings. There are many reasons why organizations are employing self-ratings. The advantages cited in the literature include the following: (a) cost effectiveness with respect to time and money (Shrauger & Osberg, 1981); (b) enhanced communication between supervisors and subordinates, which has resulted in improved relationships between supervisors and subordinates (Carroll & Schneier, 1982; Fletcher, 1986); (c) improved

communication between supervisor and subordinates, which reduces the ambiguity in the appraisal process by resolving rating disagreements (Fletcher, 1986); (d) increased ratee participation resulting in a greater acceptance of the appraisal results (Latham & Wexley, 1981; Riggio & Cole, 1992; Shrauger & Osberg, 1981); (e) reduced defensiveness in the performance evaluation (Farh et al., 1988; Latham & Wexley, 1981); (f) improved legal defensibility because of the utilization of multiple raters (Bernardin & Beatty, 1984); (g) biases in individual ratings identified by multiple ratings (Farh et al., 1988); (h) diminished halo error compared to supervisory ratings (Thornton, 1980); (i) relevant performance criteria observed (Borman, 1974; Henderson, 1984); (j) enhanced performance by the subordinate following the self-assessment (Bassett & Meyer, 1968); and (k) valuable personal information which aids supervisory ratings (Farh et al., 1988). Several researchers have also suggested that allowing subordinates to participate in the performance appraisal system has a positive effect on job satisfaction (Cleveland et al., 1989; McEnery & McEnery, 1987; Thornton, 1980). Finally, employees feel the feedback from multiple sources helps them evaluate the status of their work and provides the opportunity to analyze their abilities and competency levels (Fox & Dinur, 1988; Shrauger & Osberg, 1981).

In addition to these advantages, there has been some promising research on the validity of self-ratings. According to Cambell and Fiske (1959),

convergent validity is demonstrated when multiple sources rate a dimension similarly, whereas divergent validity is illustrated when there is independence in ratings across different dimensions. Several studies have concluded that self-ratings have at least moderate convergent validity with supervisory and peer ratings. In their meta-analysis, Harris and Schaubroeck (1988) found the mean convergence of performance ratings among 36 studies to be .35 for self-supervisory ratings and .36 for self-peer ratings. Accordingly, Somers and Birnbaum (1991) found significant correlations ranging from .27 to .41 between self- and supervisory ratings for 8 of 10 performance dimensions studied. These researchers investigated the use of a multi-trait, multi-method technique in the context of a performance appraisal.

Support for convergent validity was also found for self-supervisory ratings in a study which examined the success of military training over a two-year period (Fox & Dinur, 1988). It was concluded that self-ratings were able to predict successful training, and self-raters can adequately evaluate their own performance. Other researchers have found that when self-raters were provided with more information about the performance level of the groups, the correlation increased to .51 between self- and supervisory ratings for overall evaluation (Farh & Dobbins, 1989; Farh & Werbel, 1986). However, it should be noted that most of the previous studies have used a combination of quantifiable and

non-quantifiable dimensions for the criteria of performance appraisals. An empirical study by Schrader and Steiner (1996) examined the effects of raters using different comparative standards as their basis for quantitative performance appraisals. They concluded that when raters were provided with instructions to use more explicit standards using a quantifiable assessment, the interrater agreement coefficients ranged from .50 to .55.

Disadvantages of self-ratings. Although many advantages are cited, skepticism still remains concerning the self-evaluation method. Numerous studies have cited the disadvantages of these ratings and suggest individuals are too biased to produce reliable and valid ratings of their work performance. Despite the research previously discussed, a number of studies have found a low to moderate (i.e., .02 to .60) correlation between self- and supervisory ratings (Harris & Schaubroeck, 1988; Landy & Farr, 1980). This argument is supported in the conclusions drawn from Mabe and West's (1982) meta-analysis which found a mean validity coefficient of .29 when self-evaluations of ability were compared with objective measures of performance. Likewise, a previously mentioned study by Harris and Schaubroeck (1988), which found a correlation of .35 between self- and supervisory ratings, concluded that the mean correlation was only .22 after correcting for statistical artifacts.

Another commonly cited limitation of self-evaluations is leniency error. Leniency has been the most widely cited opposition to self-ratings (Landy &

Farr, 1980). A leniency error occurs when individuals systematically rate themselves high across all dimensions of the evaluation. Much of the literature indicates self-raters tend to inflate their appraisals and thereby threaten the validity of their ratings (Farh & Werbel, 1986; Fox & Dinur, 1988; Hoffman et al., 1991; Mabe & West, 1982; McEnery & McEnery, 1987; Thornton, 1980). Taken as a whole, these studies suggest that compared to supervisors, self-raters tend to provide inflated evaluations of their abilities and their performance. One study found that on average, self-raters felt they were performing better than three-fourths of other employees (Meyer, 1980). Although only 2 people out of the 92 participants rated themselves below the 50th percentile, they both gave themselves a rating of 45 on a 100-point scale. Further, it was concluded self-raters saw themselves as "one of the best" in terms of job performance.

On the other hand, several studies have been unsupportive of the leniency error among self-raters. Farh et al. (1988) examined the effects of a self-appraisal based performance evaluation system among faculty members and their chairpersons. The participants were asked to rate the performance of a faculty member with respect to instruction, scholarship, and departmental service. The results indicated a high agreement between the self- and chairperson ratings, and faculty member ratings were no more lenient than chairperson ratings across all performance dimensions. Another study by Shrauger and Osberg (1981) supported the lack of leniency among self-raters

when they compared self-ratings with other procedures commonly used in evaluation (e.g., peer ratings, past performance). These researchers found self-appraisals to be at least as valid as the other assessment methods utilized. This analysis was given support by a more recent study which found no leniency between self- and supervisory ratings among 106 dyads who represented nine organizations (Schrader & Steiner, 1996).

Another rater bias related to self-ratings is halo error. Cascio (1991) claimed that halo occurs when "raters have a tendency to rate an individual either high or low on many factors because the raters know (or think they know) the individual to be high or low on a specific factor" (p.84). For instance, if a rater finds a person to be physically attractive, the rater may think the person is also friendly and outgoing, generalizing from one attribute to multiple personality characteristics. However, Balzer and Sulsky (1992) maintained there is confusion concerning the operational definition of halo and that halo can actually occur in one of two forms: (a) General Impression Halo, when a rater erroneously evaluates a ratee's performance on the basis of the overall impression of the ratee or (b) Dimensional Similarity Halo, when a rater perceives different dimensions as being similar and consistently rates an individual similarly across these dimensions. Because these conceptual and operational definitions have only recently been reviewed and it is unclear as to which type of halo previous studies have examined, the conclusions drawn from

the performance appraisal literature concerning the effects of halo should be interpreted with caution.

Overall, studies have suggested there is minimal halo error in self-ratings (Fox & Dinur, 1988). Somers and Birnbaum (1991) found interrater agreement to be .64 between self- and supervisory ratings when they corrected correlations for statistical artifacts and halo error. This finding indicates that although halo was present in both sets of ratings, it did not have a significant effect on the convergence of ratings. Nonetheless, recent research has suggested that halo error may be less of a problem than previously thought. In 1990, Nathan and Tippins found that the greater the halo effect in performance appraisals, the higher the validity coefficients. This paradox alludes to the notion that halo serves to ensure that raters consider the performance of the individual as part of the whole person rather than focusing on specific behaviors which may be unrepresentative. Further, it has been proposed that the presence of halo does not necessarily indicate inaccuracy in performance ratings, rather, halo may be used as a measure of how individuals cognitively process information about other people (Balzer & Sulsky, 1992). Therefore, due to the mentioned studies which have produced conflicting results, the effects of halo need to be carefully evaluated.

Objective Versus Subjective Performance Measures and Subjective Occupations

Almost 50 years ago, Thorndike (1949) stated “the most fundamental and most difficult problem in any selection research program is to obtain satisfactory criterion measures of performance on the job against which to validate selection procedures” (p. 119). Traditionally, performance appraisals are classified as either objective or subjective in nature. An objective performance evaluation focuses on very specific, goal-driven, measurable results. On the other hand, a subjective performance appraisal is behaviorally focused and examines the quality of the work behavior. Often performance ratings in a subjective evaluation are made without pure, quantitative criteria. Subjective jobs that generally do not produce a quantifiable product are more difficult to assess because they require raters to observe behavior and then make a judgment based on the quality of performance (Clement & Stevens, 1989).

There are many occupations that are subjective in nature. The vocations of professors, firefighters, surgeons, managers, physicians, and nurses involve relatively few criteria that can be counted. Even if the criteria can be enumerated, the measures produced may not be appropriate (Berry & Houston, 1993). For example, a professor’s performance may be evaluated by counting the number of minutes they spend lecturing, the number of students they have in their classes, the number of students in their classes who receive an A, or the number of office hours they hold during a semester. Although this information

may be interesting, these measurable results are not necessarily relevant aspects of a professor's job, and this type of criteria would be inappropriate to use as the basis for performance appraisal ratings. Correspondingly, Vecchio and Gobdel (1984) proposed that approximately 9% of the variability in supervisory ratings of subordinates can be attributed to supervisors using irrelevant objective criteria. This finding suggests that although supervisors often rely on definitive criteria, these measures may not represent the most important aspects of a job.

Although the performance appraisal of faculty is difficult, much of the confusion surrounding the evaluation of this subjective occupation is the lack of set "job behaviors" that identify the excellent professor (Clement & Stevens, 1989). University administrators in Texas attempted to grade faculty's performance on the basis of a fixed scale of accomplishments, yet this practice was abandoned because it was too cumbersome and was not conducive to the rating of faculty performance (Rosenthal et al., 1994). Overall, the literature indicates that the performance of faculty members is most appropriately examined on a more qualitative basis.

However, due to this lack of definitive criteria, accurately assessing the work performance of those individuals who are in subjectively-based occupations can be difficult. Further, it is common for multiple raters to arrive at the performance appraisal process with completely different perspectives or

frames of reference (Borman, 1974; Schrader & Steiner, 1996). This problem may be compounded in traditionally subjective occupations.

Differential Comparison Standards

There are several factors that influence both objective and subjective performance evaluations. During both types of appraisal processes, raters may use different comparison standards when making their performance ratings. Schrader and Steiner (1996) found raters make comparisons with themselves, with groups, specific standards, or some combination thereof. Some research has supported the notion that the usage of differential comparison standards may be responsible for the low interrater agreement between self- and supervisory ratings. Alternatively, if the raters utilize the same comparison standard, the correlations among multiple raters increase. The study discussed earlier by Farh et al. (1988) found when faculty members and chairpersons were provided with explicit comparison standards, there was a high congruency between the ratings. Likewise, two additional reviews of the self-rating validity concluded interrater agreement can be increased when raters are provided with instructions on who to use as social comparisons (Heneman, 1986; Mabe & West, 1982).

Previous research has investigated the referent groups people select when evaluating their situation. Goodman (1974) identified three potential referent sources. First, individuals may make comparisons with themselves

regarding their inputs and outputs. Second, people may compare themselves to “others.” This “others” group could include other individuals inside the organization or other people outside the work setting. Third, people may select the system as a referent. By comparing themselves to the system, individuals are comparing themselves with the aspects of the pay system and the administration of their organization. Summers and DeNisi (1990) expanded Goodman’s list to nine referents and concluded it is important to indicate referent selections when evaluating social comparisons. In general, employees tended to select multiple reference standards. In the study mentioned earlier, Schrader and Steiner (1996) also found self-raters and supervisors tend to prefer multiple comparison standards. The equity theory and the social comparison theory are two important perspectives to examine when determining which of these comparison standards individuals utilize.

Equity theory. Adam’s (1965) theory of equity states individuals form a ratio of their inputs (e.g., performance on the job, work experience, training) in a situation to their outcomes (e.g., salary, benefits, job security) in that situation (as cited in Cascio, 1991). An input is defined as anything the person feels he or she is contributing to the situation, and an outcome is the reward the person is receiving in return. An individual will compare the value of this input/outcome ratio with other people. This comparison group can consist of co-workers inside the organization or individuals in a similar job who are outside the company.

Equity exists when individuals perceive their own ratio to be equal to those of other people. This perceived equity will result in satisfaction for the individual. However, as Adams suggests, tension will arise if a person perceives the input/output ratio as inequitable relative to others.

The study mentioned earlier by Summers and DeNisi (1991) reexamined Goodman's (1974) study of referent selections by using equity theory. They explored the degree of perceived inequity among managers with regards to compensation. Their results indicated 34% of the managers used self as their referent, 20% used others who were within their organization, almost 6% used individuals who were outside their organization as their referent of choice, and over 37% used some form of these standards to make their comparison.

Another study that evaluated the selection of referents among individuals who had perceptions of pay inequity also found co-workers to be a common referent choice (Dornstein, 1989). These findings suggest equity theory may help explain the choice of referents among individuals.

Social comparison theory. Another psychological theory related to the selection of referents is Festinger's (1954) social comparison theory. Festinger claimed that because people desire stable and accurate assessments of their behavior, individuals compare their abilities with other people. In the work environment, employees may compare themselves with their co-workers to develop these perceptions of behavior. Research indicates that individuals

select referents who are more similar to themselves in traits and abilities (Baron & Graziano, 1991). In the context of a performance appraisal for subjective occupations, people may be more prone to this social comparison because there are no objective measures to use as a guideline for determining successful performance. Rather, it is a qualitative judgment that is behaviorally based. Therefore, the critical question becomes with whom do individuals compare themselves when making such judgments.

It appears self-raters tend to use comparisons which will make themselves look good. Fisher (1989) found self-raters' referent choice may be systematically biased so they are comparing themselves to someone who is perceived as having less ability. This comparison may lead to a positive, but inaccurate rating of one's performance. This type of comparison may help explain the tendency for self-raters to be more lenient in their ratings. Although supervisors generally do not succumb to this inaccurate leniency bias, when they are rating an employee, they will often have a different perspective of the employee's performance (Schrader & Steiner, 1996). Consequently, if the self-rater is using one comparison standard and the supervisor is using a different comparison standard, it is very likely the interrater agreement of performance will be low.

Classification of comparison standards. In his original theoretical framework, Adams (1965) indicated the most common type of referent choice is

another individual (“other”) (as cited in Cascio, 1991). Then researchers began to suggest individuals will compare themselves to other people within the particular organization as well as individuals who work outside the unit (Goodman, 1974; Oldham, Kulik, Ambrose, Stepina, & Brand, 1986). In a recent examination of personal and situational variables involved in the selection of referents, Kulik and Ambrose (1992) indicated employees prefer to use personal, internal standards as their referent. This study was an expansion of the conclusions drawn from a previous similar study that found when people are given a choice, they will most likely select their own personal values as a referent. Specifically, the participants in the latter study selected themselves as a referent over 56% of the time (Oldham et al., 1986).

As discussed earlier, a contemporary study by Schrader and Steiner (1996) investigated differential comparison standards in a performance appraisal framework. These researchers demonstrated that employees have access to four different comparison standards: (a) internal (i.e., a comparison to oneself), (b) absolute (i.e., a comparison to some objective measure), (c) relative (i.e., a comparison to others), and (d) multiple (i.e., a comparison utilizing internal, absolute, and relative standards). However, in their study, these researchers did not distinguish whether an individual made comparisons to individuals inside the particular organization or outside the organization. To determine exactly who employees use when they make comparisons requires dividing this relative

standard into two subparts: (a) relative-inside (i.e., a comparison to others working within the organization) and (b) relative-outside (i.e., a comparison to others outside the organization). Goodman's (1974) typology did include this differentiation but did not include the possibility of having the multiple comparison standard. Therefore, the performance appraisal literature is lacking a comprehensive investigation of the differential comparison standards. One of the purposes of the present study was to examine the use of the following six comparison standards: ambiguous, internal, absolute, relative-inside, relative-outside, and multiple.

The second purpose of this study was to investigate how the use of these six differential comparison standards will be used when they are applied to subjective occupations. The empirical study by Schrader and Steiner (1996) evaluated the use of differential comparison standards by occupations that already used quantifiable criteria for performance evaluations. In the present study, the researcher sought to explain the use of six differential comparison standards when applied to the performance evaluations of occupations which are traditionally subjective in nature.

Hypotheses of the present study

Based on the literature discussed, the following hypotheses were proposed:

Hypothesis 1. Both self- and supervisory ratings will significantly differ as a function of the comparison standards (ambiguous, internal, absolute, relative-inside, relative-outside, and multiple) utilized in the rating instructions.

Hypothesis 2. The interrater agreement between self- and supervisory ratings, when collapsed across the three performance dimensions, will be significantly higher for the explicit comparison standards (internal, absolute, relative-inside, relative-outside, multiple) than for the ambiguous comparison standard (which does not provide the rater with any explicit instructions as to which comparison standard to use).

Hypotheses 1 and 2 are essentially a replication of Schrader and Steiner's (1996) hypotheses. However, the present study applied these hypotheses to a different sample from a subjectively-based occupation and include an additional comparison standard.

The remaining hypotheses examined which of the comparative standards (ambiguous, internal, absolute, relative-inside, relative-outside, or multiple) the raters previously utilized and which of the standards they preferred to use in future performance ratings.

Hypothesis 3. When supervisors are asked which of the six comparison standards they have used in the past to decide whether or not a faculty member was performing satisfactorily on the job, they will indicate they have used the relative-outside standard most often.

Hypothesis 3 is related to a previous study which found that individuals with greater levels of professionalism tend to select referents outside their organization. Individuals in professional occupational roles have more interorganizational mobility and have more access to information about outside referents (Goodman, 1974). Therefore, these individuals consider outside referents as appropriate benchmarks to use when evaluating a subordinate's work performance.

Hypothesis 4a. Due to the subjective nature of the participant's occupation, both self-raters and supervisory raters will have the least preference for the absolute comparison standard when asked which one they would select to use in future performance ratings .

Hypothesis 4b. Due to the subjective nature of the participant's occupation, both the self-raters and supervisory raters will prefer the multiple comparison standard when asked which one they would select to use in future performance ratings.

Hypothesis 4c. Due to the subjective nature of the participant's occupation, both the self-raters and supervisory raters will prefer the internal standard (after the multiple standard), followed by the relative-outside, relative-inside, and absolute respectively, when asked which one they would select to use in future performance ratings.

Hypothesis 4a was based on the rationale that because

subjectively-based occupations produce relatively few quantifiable products, the work performance of individuals in these occupations is behaviorally focused and is evaluated using qualitative judgments (Clement & Stevens, 1989; Rosenthal et al., 1994). Heneman (1986) and Mabe and West (1982) suggested that self-raters object to definitive measures of performance. Hypothesis 4b was based on the conclusions drawn from Schrader and Steiner's (1996) study which found that individuals prefer to use a multiple comparison standard, that utilizes the most comprehensive information, when making performance ratings. Hypothesis 4c was linked to the findings of Kulik and Ambrose (1992) and Oldham et al. (1986) which suggested individuals prefer to use an internal comparison standard. The hypothesized sequential order of preferences also stemmed from Festinger's (1954) social comparison theory which asserts that in the absence of absolute standards, individuals will compare themselves to others to form an accurate assessment of behavior. Likewise, as suggested in Mabe and West's (1982) meta-analysis, performance evaluations may be more of a contrasting measurement process rather than an absolute one, therefore, a benchmark based on comparisons may be favored over the absolute standard.

CHAPTER 2

METHOD

The purpose of this study was to determine if interrater agreement between self- and supervisory ratings increased when both raters were assessing the performance of individuals in subjective occupations and were provided with similar comparison standards. The following section discusses the procedures that were implemented in this investigation. The dyads in the sample consisted of tenured faculty members and their chairpersons. It was argued the position of a faculty member is a subjective occupation. Because subjective occupations are behaviorally focused, raters must observe a worker's behavior and make a judgment based on the quality of work performance. The vocation of a faculty member involves relatively few quantifiable criteria and is clearly an example of a subjective occupation (Rosenthal et al., 1994). The qualitative nature of this occupation provided an ideal opportunity to examine the use of differential comparison standards in subjective performance appraisals. Data was collected by mailing packets to the participants.

Participants

The participants for this study included chairpersons and faculty members who were selected among six midwestern universities. All of the divisional chairpersons at each institution were selected to participate in the research. Only faculty members who had tenure, were working full-time, and had been

employed at their current job for at least two years were used in the study. In addition, the faculty members had to be supervised and evaluated by the chairperson. This criterion was necessary in order to examine the comparative referent group for the relative-inside comparison standard.

After obtaining a list of all of the chairpersons from the respective administration offices, packets containing a survey for the chairperson and a survey for one of his/her faculty were mailed to each of the chairpersons. Specifically, a total of 200 packets (400 surveys) were sent to the chairpersons. Upon receiving the packets, the department chairpersons selected the faculty member he or she evaluated. Of the 400 surveys distributed (200 chairpersons and 200 faculty members), 100 surveys were completed and returned. Therefore, a response rate of 25% was obtained. However, 10 of these were invalid because either they were incomplete ($n = 3$), one member of the supervisor-faculty member dyad did not return the survey ($n = 4$), or the respondent did not understand the survey ($n = 3$). A total of 90 surveys (45 chairpersons and 45 faculty members) were used for the analysis.

The participants' age ranged from 31 to 65 years of age with a mean of 49.83. Eighty-three percent were men, and 17% were women; the average number of faculty members under the each chairperson's direct supervision was 13.3. For faculty members, the average number of years at their present job

was 11.4, whereas chairpersons had worked in their position for an average of 16.7 years.

Procedure

After obtaining permission from each of the university's research committees and the Institutional Review Board for Treatment of Human Subjects of Emporia State University (Appendix A), the data collection began. The initial step involved the researcher contacting each chairperson to obtain a list of the faculty members who also participated.

An envelope was mailed to each of the chairpersons. Inside the envelope, a survey for each of the chairpersons and faculty members included a transmittal letter, an informed consent document, a series of performance appraisal rating sheets, rating-related questions, a demographic profile, and a self-addressed stamped envelope. The transmittal letter (Appendix B) explained the purpose of the study, the importance of their participation, and the confidentiality of participants. The informed consent document (Appendix C) provided detailed instructions for the survey and indicated that participation in the study was completely voluntary.

Measures

Self-evaluations. The faculty members were asked to rate themselves on three performance dimensions: Scholarship, Service, and Teaching/Instruction. A review of the duties and obligations of professors indicated these three main

components represent 76% of faculty workload (Rosenthal et al., 1994). Hence, most of the studies concerning faculty members' work performance have used these three dimensions as the criteria for faculty evaluations (Clement & Stevens, 1989; Farh et al., 1988). In the present study, Scholarship represented the most objective criteria of the three dimensions as it was based on the number of publications printed in a professional journal, books or chapters. Service included participation and involvement in distinct university or department committees the faculty member has served on at his/her university. Teaching/Instruction, the most subjective dimension, consisted of students' evaluations on the faculty member's teaching methods/techniques. These performance dimensions were defined for all of the participants (Appendix D). The faculty members were asked to rate their performance on all dimensions as a college professor over the past two years. Although these three dimensions are considered to be the main activities of faculty, the amount of time devoted to each varies (Rosenthal et al., 1994). Based on this rationale, the participants were asked to consider their performance over the two-year time span rather than just one year.

The three dimensions were based on a 9-point graphic rating scale (1 = Very Poor, 3 = Poor, 5 = Average, 7 = Good, and 9 = Very Good). All faculty members completed six rating sheets. Each sheet of paper provided the participant with instructions to use different comparison standards (Appendix E).

The faculty members and chairpersons were provided with instructions asking them to use the following six comparison standards: (a) **AMBIGUOUS** - "Based on your performance over the past year, please rate yourself on the following dimensions.", (b) **INTERNAL** - "Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use your own personal, internal values and standards as a criteria.", (c) **ABSOLUTE** - "Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use the average requirement or goal listed in the parentheses next to each dimension as a criteria.", (d) **RELATIVE-INSIDE** - "Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use the performance of fellow faculty members who work within your university as a criteria.", (e) **RELATIVE-OUTSIDE** - Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use the performance of other individuals who have similar jobs and who work in comparable departments, but work outside your university as a criteria.", and (f) **MULTIPLE** - "Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use your own personal standards, your attainment of the average requirement and goals, and your comparison with other faculty members both within and outside your university as a criteria."

Below the instructions on the absolute comparison standard rating sheet, the raters were provided with the average performance of a faculty member over a two-year period. First, the average number of publications in a two-year period, based on the results from a 1995-1996 survey of 34,000 professors, is three (Magner, 1996). Therefore, the rating sheets indicated this average for the scholarship performance dimension. Second, the objective measure for the service dimension, the number of distinct committees served on, was four. Third, the measurable criteria for the instruction dimension was linked to the average student ratings of faculty members at a medium-sized, midwestern university, which is 4.0 on a 5-point scale.

With the exception of the ambiguous comparison standard rating sheet, all of the sheets were titled according to their appropriate standard. The ambiguous rating sheet was always presented first because it represented an undefined standard, and the multiple rating sheet was always presented last because it represented a combination of internal, absolute, relative-inside, and relative-outside comparison standards. However, the internal, absolute, relative-inside, and relative-outside rating sheets were introduced randomly so that the researcher could control for order effects.

Chairperson evaluations. Each chairperson was presented with similar rating sheets and asked to rate the faculty member on all of the same three performance dimensions. The chairperson was asked to use the same graphic

rating scale and the same comparison standards. The rating sheets were presented in the same randomized fashion as the faculty member sheets. However, wording for the instructions on the rating sheets was altered slightly for the chairpersons (Appendix F).

Post-rating comparison standard ratings. In order to determine which comparison standard the participants have used in the past or would prefer to use in future performance evaluations, the participants were asked to respond to post-rating questions (Appendix G). The chairpersons were asked, "Please think about how you have rated you faculty member's job performance prior to answering this packet. Based on the previous five comparison standards (internal, absolute, relative-inside, relative-outside, and multiple), which one have you used most often in the past as the basis for you ratings?" The chairpersons were prompted to circle the appropriate comparison standard.

To determine which of the comparison standards both of the raters would prefer to use in the future, all participants were asked to respond to one final question about the comparison standards. The chairpersons were questioned, "If asked to rate a faculty member's performance in the future, please rate each of the five comparison standards as to your preference for using them in future performance appraisals." The faculty members were asked a similar question, "If asked to evaluate your own performance in the future (i.e., provide a

self-rating), please rate each of the five comparison standards as to your preference for using them in future performance ratings.” The chairpersons and faculty members were asked to rate their preference on a 9-point rating scale.

Demographics and comprehension question. A demographic profile was also included in all of the participants’ packet. The demographic form (Appendix H) consisted of items relating to sex, age, current job title, and tenure with their university. In addition, chairpersons were asked how many faculty members are under their direct supervision.

At the bottom of the demographic profile, the researcher asked the participants to answer a question regarding their understanding of the appraisal process. The participants were asked, “Do you feel you understood all the instructions and questions asked throughout this packet and were able to answer them in an honest and accurate manner?.” The responses to this question allowed the researcher to determine whether the participants comprehended the instructions of the study.

CHAPTER 3

RESULTS

The primary purpose of this study was to determine if the interrater agreement between self- and supervisory performance ratings would increase when both raters were provided with similar comparison standards and were assessing the performance of individuals in subjective occupations. The participants in the study included 90 tenured university professors. Specifically, the sample contained 45 department faculty member-chairperson dyads.

Hypothesis 1

The first hypothesis investigated the differential effects of the comparison standards on performance ratings. It was hypothesized, depending upon the comparison standard utilized, there would be significant differences among the ratings. This hypothesis was tested using a 2 x 3 x 6 (rater source x performance dimension x comparison standard) repeated measures analysis of variance (ANOVA), and the eta squared statistic was used to calculate the effect size of the significant effects. The within-subject variables were performance dimension and comparison standard, whereas rater source represented the between-subject variable. The results, as illustrated in Table 1, yielded a nonsignificant effect for rater source, $F(1, 88) = .40, ns$. This result suggests there were no significant differences in ratings between the raters (faculty member, chairperson) and suggests a lack of leniency among the faculty

Table 1

Repeated Measures Analysis of Variance for Performance Ratings:
Rater Source x Performance Dimension x Comparison Standard

| SOURCE | <u>SS</u> | <u>df</u> | <u>MS</u> | <u>F</u> |
|---------------|-----------|-----------|-----------|----------|
| RATER (R) | 5.57 | 1 | 5.57 | .40 |
| Error | 1216.48 | 88 | 13.82 | |
| DIMENSION (D) | 401.80 | 2 | 200.90 | 25.56* |
| R x D | 9.50 | 2 | 4.75 | .60 |
| S x D | 37.47 | 10 | 3.75 | 7.76* |
| Error | 383.14 | 176 | 7.86 | |
| STANDARD (S) | 58.64 | 5 | 11.73 | 11.96* |
| R x S | 5.91 | 5 | 1.18 | 1.20 |
| Error | 433.06 | 440 | .98 | |
| R x S x D | 3.93 | 10 | .39 | .81 |
| Error | 424.83 | 880 | .48 | |

Note. N = 90

* $p < .001$

member ratings. However, a significant main effect for performance dimensions was attained, $F(2, 176) = 25.56, p < .001, \eta^2 = .23$, meaning both the chairpersons and faculty members were rating each of the performance dimensions (scholarship, service, teaching/instruction) differently. A main effect for comparison standard was also significant, $F(5, 440) = 11.96, p < .001, \eta^2 = .12$. This finding suggests the participants' ratings were different depending on which comparison standard (ambiguous, internal, absolute, relative-inside, relative-outside, multiple) they were using. The presence of a significant comparison standard by dimension interaction, $F(10, 176) = 7.76, p < .001$, indicated ratings on the different performance dimensions must be considered in light of which comparison standard the raters were using. As illustrated in Figure 1, the ratings on Performance Dimensions 2 and 3 (service and teaching/instruction) were higher than those on Performance Dimension 1 (scholarship), and ratings were consistently lower on all three dimensions when raters were using the absolute comparison standard. Finally, with the exception of Performance Dimension 1 using the absolute standard, most of the ratings were within a 2-point range. Due to the significant interaction, two separate one-way ANOVAs were computed for each of the significant main effects. These calculations also produced significant results for both comparison standard,

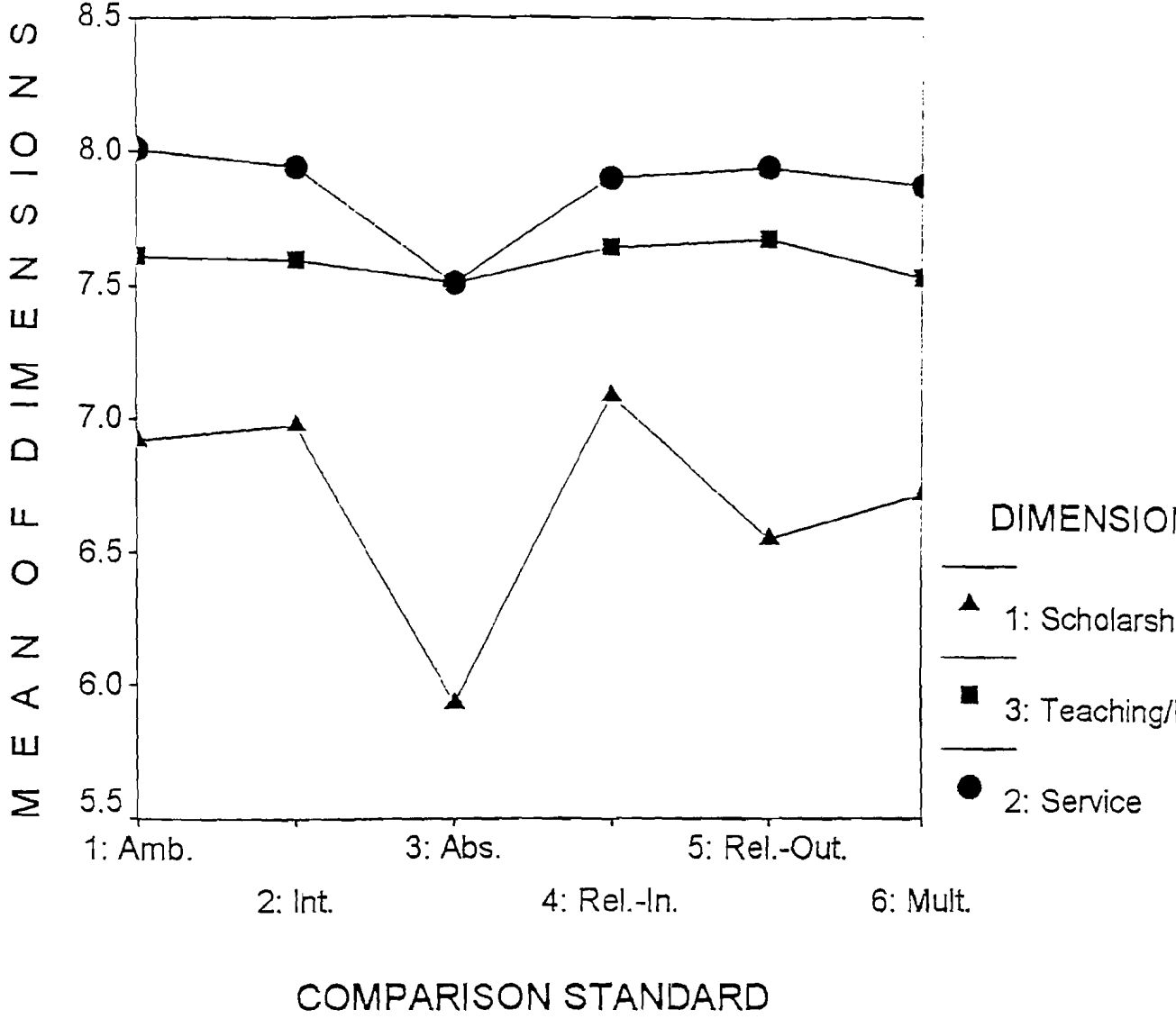


Figure 1. Comparison Standard x Performance Dimension Interaction

$F(5, 84) = 7.15, p < .001, \eta^2 = .30$, and performance dimension, $F(2, 87) = 20.20, p < .001, \eta^2 = .32$, indicating further support for the effect of these two variables as they were significant when analyzed individually.

A Tukey's honestly significant difference (HSD) multiple comparison procedure was used to identify specific group mean differences between the comparison standards. The Tukey's findings demonstrated that ratings associated with the absolute comparison standard were significantly lower than ratings made with all of the other five standards (see Table 2). Average ratings from the ambiguous, internal, and relative-inside comparison standards were not significantly different from each other, but did approach significance when compared to the absolute, relative-outside, and multiple comparison standard. Overall, the raters produced the highest ratings when they used the relative-inside comparison standard.

The combined results of the three-way ANOVA, separate one-way ANOVAs, and Tukey's procedures were supportive of Hypothesis 1. The statements associated with this hypothesis predicted both the chairperson and the faculty member ratings would significantly differ as a function of the six comparison standards, and the findings clearly demonstrated the significance of these standards.

Table 2

Tukey's HSD Analysis of Differences Between Comparison Standard Means For Full Sample

| Comparison Standard | Mean |
|---------------------|----------------------------|
| RELATIVE-INSIDE | 7.54 ^a (.10) |
| AMBIGUOUS | 7.52 ^{a, b} (.11) |
| INTERNAL | 7.51 ^{a, c} (.10) |
| RELATIVE-OUTSIDE | 7.39 ^{b, c} (.11) |
| MULTIPLE | 7.38 ^c (.10) |
| ABSOLUTE | 6.99 ^d (.13) |

Note. N = 90. Means with different superscripts are significantly different ($p < .05$). Standard deviations in parentheses.

Hypothesis 2

Hypothesis 2 stated that the interrater agreement between self- and supervisory ratings, when collapsed across the performance dimensions, would be significantly higher for the explicit standards than for the ambiguous standard. This hypothesis was analyzed using the Pearson product-moment correlational method. Specifically, this hypothesis was tested by using a two-sample correlational t -test for independent samples. The means for faculty member and chairperson ratings across the different comparison formats were calculated. These six faculty member-chairperson correlations for each comparison standard were then tested against one another to see if they significantly differed. Table 3 reports the mean correlations between the two raters for each comparison standard. The mean correlations for each of the standards were not found to be significantly different from each other, ranging from .34 to .15. Apparently, the level of interrater agreement did not significantly change when raters used differential comparison standards. This finding suggests the various comparison standards did not significantly affect the interrater agreement between chairperson and faculty member performance ratings. However, further analysis of these correlations indicated when both raters used the ambiguous, internal, absolute or multiple comparison standards, the correlation coefficients were statistically greater than zero. The conclusion

Table 3

Pearson Correlations Between Faculty Member Ratings and Chairperson Ratings for Each Comparison Standard

| Comparison Standard | r |
|---------------------|------------------|
| INTERNAL | .34 ^a |
| ABSOLUTE | .34 ^a |
| AMBIGUOUS | .33 ^a |
| MULTIPLE | .27 ^a |
| RELATIVE-INSIDE | .16 |
| RELATIVE-OUTSIDE | .15 |

Note. $n = 45$. Superscript indicates correlation is statistically significant ($p < .05$).

indicates there is a relationship between the comparison standards and the interrater agreement, but this conclusion should not be construed to imply the specific magnitude of the relationship, except that it is simply significantly different from zero.

Hypothesis 3

The third hypothesis predicted the chairpersons would report they have used the relative-outside comparison standard most often in previous performance appraisals. Following a frequency count of the responses, a chi-square analysis was used to further investigate the usage of the comparison standards. The chi-square results were significant, $\chi^2(4, N = 45) = 19.33, p < .001$, and supportive of Hypothesis 3 as it had been anticipated chairpersons would proclaim using the relative-outside comparison standards as the basis for previous performance appraisals of their subordinates. The expected and observed frequencies are shown in Table 4. Forty percent of the chairpersons indicated using the relative-outside standard most often, and 24.4% reported utilizing the relative-inside and multiple comparison standard. However, only 6.7% and 4.5% declared using the absolute and internal comparison standard, respectively.

Hypothesis 4

Hypothesis 4 predicted faculty members and chairpersons would prefer to use the multiple standard, followed by the internal, relative-outside,

Table 4

Observed and Expected Frequencies for Previously Used
Comparison Standards Reported by Chairpersons

| COMPARISON STANDARD | OBSERVED N | EXPECTED N | OBSERVED % |
|------------------------|---------------|---------------|---------------|
| INTERNAL | 9 | 2 | 4.5 |
| ABSOLUTE | 9 | 3 | 6.7 |
| RELATIVE-INSIDE | 9 | 11 | 24.4 |
| RELATIVE-OUTSIDE | 9 | 18 | 40.0 |
| MULTIPLE | 9 | 11 | 24.4 |

Note. N = 45

relative-inside, and absolute respectively, when asked which comparison standard they would prefer to use in future performance ratings. This hypothesis was assessed by comparing the mean differences in the preference ratings. Table 5 shows the means and standard deviations for the faculty member and chairperson preference ratings.

Hypothesis 4a predicted both raters would have the lowest preference for the absolute standard. Although the results indicated the relative-outside standard as the least preferred, the ratings for the absolute comparison standard were in the lower half of the faculty member's and chairperson's ratings as well as the full sample's ratings. Nonetheless, Hypothesis 4a was not fully supported. Hypothesis 4b was confirmed as the total average for both chairpersons and faculty members indicated a preference towards using the multiple comparison standard. Hypothesis 4c received little support because it was predicted both raters would prefer the internal (following the multiple standard), then the relative-outside, relative-inside, and absolute, respectively. However, the results in this study indicated the rater's order of preference for the comparison standards were follows: multiple, relative-inside, internal, absolute, relative-outside for the full sample. Of particular note is the least preferred comparison standard, relative-outside, as Hypothesis 3 found this referent to be the most commonly used by chairpersons. Thus, although chairpersons claimed to have used the relative-outside comparison standard most often, they also

Table 5

Means and Standard Deviations for Faculty Member-Chairperson Preference Ratings

| COMPARISON STANDARD | FACULTY MEMBER | CHAIRPERSON | FULL SAMPLE |
|---------------------|----------------|----------------|----------------|
| MULTIPLE | 6.58 (1.70) | 6.98 (1.70) | 6.78 (1.70) |
| RELATIVE-INSIDE | 6.11 (2.05) | 7.16 (1.40) | 6.63 (1.82) |
| INTERNAL | 6.84 (1.91) | 6.00 (1.87) | 6.42 (1.93) |
| ABSOLUTE | 5.67 (2.20) | 6.09 (1.73) | 5.88 (1.98) |
| RELATIVE-OUTSIDE | 5.64 (1.93) | 5.24 (1.84) | 5.44 (1.88) |

Note. $n = 45$ for faculty member and chairperson; $N = 90$ for full sample. Standard deviations in parentheses.

reported this referent to be the least preferred. Collectively, these results provided mixed support for Hypothesis 4.

CHAPTER 4

DISCUSSION

The present study was conducted to examine the effects of differential comparison standards on the level of interrater agreement between self- and supervisory performance ratings. Six comparison standards (ambiguous, internal, absolute, relative-inside, relative-outside, and multiple) were applied to performance evaluations of an occupation which is traditionally subjective in nature. The participants were tenured university chairpersons and faculty members.

Overall, the results supported a main effect for the comparison standards and the different performance dimensions (scholarship, service, and teaching/instruction). Therefore, the raters' ratings were clearly dependent upon which comparison standard format they were utilizing and which performance dimension they were evaluating. Also, as anticipated, the findings demonstrated when chairpersons were asked which of the five explicit comparison standards they had used the most in previous performance evaluations, they reported using the relative-outside referent most often. Furthermore, the raters tended to prefer the multiple comparison standard for future appraisals, yet the specific trend of preference ratings varied among the rater source. Only one hypothesis yielded less than supportive evidence as the interrater agreement between

faculty member and chairperson ratings did not significantly increase when both raters used the same comparison standard.

Interpretation of Results

The first hypothesis predicted the faculty member and chairperson ratings would differ as a function of which comparison standard they were using. The results strongly supported this hypothesis and the existence of differential comparison standards. That is, faculty member's and chairperson's ratings were significantly different depending upon which comparison standard was utilized. This finding is supportive of Schrader and Steiner's (1996) study on the effects of differential comparison standards. Overall, the ratings using the ambiguous, internal, and relative-inside standards were not significantly different, yet they did achieve significance when compared to the absolute, relative-outside, and multiple comparison standards. The relative-inside standard produced the highest performance ratings, and the absolute comparison standard, the most objective referent, yielded statistically lower ratings than the other five comparison standards.

The significant main effect of performance dimensions suggests the raters in this study viewed the performance dimensions (scholarship, service, and teaching/instruction) as three separate responsibilities of a college faculty member. The highest ratings were on the service dimension, and scholarship received the lowest ratings. However, all mean ratings for the performance

dimensions were higher than 5.5 on a 9-point scale, suggesting faculty members were rated above average on most performance dimensions.

The significant comparison standard x performance dimension interaction indicates the raters' evaluations were dependent upon the comparison standard they were instructed to use and the performance dimension they were considering. To further investigate this hypothesis, two separate one-way ANOVAs were performed producing significance for performance dimension and comparison standard. These analyses add additional credence to the significant effects of different comparison standards and the various performance dimensions on performance ratings. This finding may also help explain the disagreement among multiple raters. That is, the use of these different comparison standards and different performance dimensions could be the underlying mechanism in the traditionally low interrater agreement between self- and supervisory performance ratings.

No rater differences were found across the different comparison standards demonstrating a lack of leniency among the self-raters. This finding supports previous studies which found no leniency among self-raters in many occupations (Schrader & Steiner, 1996; Somers & Birnbaum, 1991), including college faculty positions (Farh et al., 1988). One possible explanation for this lack of leniency may be that the faculty members tended to be more accurate in their ratings because they expected their ratings to be compared with their

superior's ratings. However, the scarcity of leniency among self-raters is incongruent with prior research that suggests self-evaluations are more susceptible to leniency bias when they are using ambiguous measures (Farh & Werbel, 1986).

The results of Hypothesis 2 rendered the least significant evidence in this study. It was hoped the interrater agreement between faculty member and chairperson ratings would significantly increase when both raters were utilizing similar comparison standards. However, the interrater agreement was consistent, or in some cases less, than the self-supervisory correlation of .35 found in Harris and Schaubroeck's (1988) meta-analysis.

The results of the current study are incongruent with the Schrader and Steiner's (1996) study which found a mean correlation between self- and supervisory performance ratings to be .55. Nonetheless, it is important to note these researchers used a sample of individuals whose work performance was based on substantive, definitive measures. In this type of objective evaluation, it is evident if the employee is working successfully because the rater has a precise criterion to use. In defense of the results in the present study, it is argued that the type of occupation had a moderating effect on the interrater agreement. That is, the subjective nature of a college professor's work performance may have strongly impacted the level of agreement between raters.

As the research clearly indicates, occupations that do not produce quantifiable products are very difficult to evaluate (Clement & Stephens, 1989).

The third hypothesis produced significant and supportive results as the chairpersons reported they have used the relative-outside comparison standard most often in previous performance appraisals of their subordinates. This is congruent with Goodman's (1974) study, which suggests the greater level of education and professionalism, the more likely one will select a referent outside his or her focal organization. Individuals in highly professional positions have greater access to these outside referent benchmarks. For example, college professors may have more interorganizational mobility as many of them attend professional conferences and meetings. Thus, the interaction with educators from different institutions may influence who they compare themselves with when judging the effectiveness of their own performance. However, this can be problematic in performance appraisals as multiple raters may be using different outside referents. Thus, the self-rater may be basing his or her rating on members of a professional organization while the supervisor is using individuals at different institutions, or in this case, different universities. The low interrater agreement found in Hypothesis 2 exemplifies this particular issue.

The results from Hypothesis 4 produced mixed support for the predicted trend in rater preferences for the differential comparison standards. Although the absolute comparison standard did not receive high ratings of preference, it

was not rated as the least preferred. Instead, both the chairpersons and the faculty members had the least preference for the relative-outside comparison standard. This finding is somewhat unsupportive of Hypothesis 4a and previous studies that have suggested self-raters having the lowest preference for objective measures of performance (Heneman, 1986; Mabe & West, 1982). However, when examining faculty member's preference ratings, the difference between averages for relative-outside and absolute standards was minimal, with means of 5.67 and 5.64, respectively. Therefore, as the literature indicates, self-raters do tend to dislike definitive performance measures.

The low preference ratings for the absolute standard also emphasizes the subjective nature of the participant's occupation. Schrader and Steiner (1996) found the absolute comparison standard to be one of the most preferred referents utilized, yet these researchers examined the different comparison standards within the context of occupations that were primarily based on objective measures. The self-raters in the prior study also rated the internal standard as one of the least preferred, whereas in the present study, the self-rater preferred the internal comparison standard most. This finding reiterates the subjective performance appraisal system in the current study as self-raters preferred to utilize a more subjective referent.

Overall, the multiple comparison standard was selected to be the most preferred comparison standard to use in future performance evaluations.

Apparently, raters prefer to use the all-inclusive referent, which is comprised of the internal, absolute, relative-inside, and relative-outside comparison standard. Hypothesis 4b was confirmed and is supportive of previous research (Kulik & Ambrose, 1992; Oldham et al., 1986; Schrader & Steiner, 1996). Nonetheless, when examining different rater's responses, the chairpersons preferred the relative-inside standard, whereas the faculty members preferred the internal comparison standard.

Lastly, Hypothesis 4c received little support as it was anticipated the raters would prefer the multiple, internal, relative-outside, relative-inside, and absolute, respectively. However, the results revealed a different trend in preference ratings and overall, the raters preferred the multiple, relative-inside, internal, absolute, and then the relative-outside comparison standard. It is particularly interesting to note that chairpersons had the lowest preference for the relative-outside referent in future performance appraisals, yet they reported using this comparison standard most often in previous evaluations. Perhaps after being exposed to the different comparative standards used in the study, some of them realized the standard they were previously using was not the most effective one. Rather, many may have found that considering more complete information, as the comprehensive multiple standard does, is better. It may be that due to the lack of definitive criteria in subjective occupations, this inclusive

multiple comparison standard is the most appropriate referent to use in academic performance appraisals.

It is clear Adam's (1965) equity theory and Festinger's (1954) social comparison theory were indirectly supported in this study. Both the chairpersons and the faculty members did make comparisons with themselves, with groups, with specific standards, or a combination of these referents. The two theories are important perspectives involved in determining which of these comparison standards individuals utilize.

Limitations

Despite some interesting findings, there are some limitations to the present study. First, the sample size was relatively small and homogeneous ($N = 45$ dyads). Due to practical restraints and accessibility, only professors from six midwestern universities were asked to participate in the study. The characteristics of the sample should be taken into consideration when generalizing the results of this study.

A second limitation centers around the response rate. As with all mail-out surveys, it is very difficult to obtain a high response rate. On the due date of the surveys, the researcher telephoned each of the chairpersons who had not responded to the study. A final follow-up call was also made to each of the individuals who represented an incomplete faculty member-chairperson dyad.

Although these procedures prompted several participants to return the packets, a low response rate of 25% was still obtained.

A third limitation to this study may have involved the participant's misunderstanding of the comparison standards, performance dimensions, or any other part of the survey. However, only 3 out of 100 respondents answered "No" to the question on the survey asking them if they understood all of the instructions and questions in the survey. Although these individuals were excluded from the study, comprehension of the survey may have produced a confound.

A final limitation to the current study is related to the subjective nature of college professor's performance evaluations. As stated, it is difficult to appraise the performance of these individuals due to the lack of definitive measures. Therefore, the subjective occupation may have been a moderator of the obtained results.

Implications and Future Research

It is important for all companies to monitor the effectiveness of their internal procedures. As employers attempt to improve their organizational processes, most have included self-ratings in their performance appraisal system or are considering such an alternative. There is a plethora of research suggesting numerous benefits of self-evaluations, yet many individuals remain

cynical about the use of these ratings (Cascio, 1991). One of the reasons for this skepticism is the lack of convergence between self- and supervisory ratings.

Performance appraisals of subjectively based occupations have not received much attention in the literature to date. It is difficult to evaluate the work performance of individuals when they generally do not produce quantifiable products. Rather, these appraisals must be made on the basis of qualitative judgments. This lack of definitive criteria may further prompt multiple raters to approach the performance evaluation with different perspectives. The aforementioned limitations notwithstanding, the results of this study have demonstrated that multiple raters do approach the performance evaluation with different perspectives and use different benchmarks for their ratings. It is hoped this study will help organizations carefully evaluate their current performance appraisal approaches and attempt to understand the disagreement among multiple raters. The current performance evaluation design may be used as a training tool to provide raters with a specific frame of reference to use as a benchmark for their ratings.

Further research is needed to examine the differential effects of the comparison standards when applied to different occupations requiring subjectively based performance appraisals. For example, the results of the present study could be explored with the occupations of managers, firefighters,

or physicians. This investigation may help explain the moderating effects of objective versus subjective occupations.

As an example of another application, the performance appraisal format could be applied to a larger sample of participants in diverse geographic locations. Only with more research will we know the full benefits and limitations of using this type of performance appraisal format.

REFERENCES

- Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. Journal of Applied Psychology, 21, 421-430.
- Baron, R. M., & Graziano, W. G. (1991). Social psychology. Orlando, FL: Holt, Rinehart and Winston.
- Bassett, G. A., & Meyer, H. H. (1968). Performance appraisal based on self-review. Personnel Psychology, 21, 421-430.
- Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston: Kent Publishing Company.
- Berry, L. M., & Houston, J. P. (1993). Psychology at work. Dubuque, IA: William C. Brown Communications.
- Borman, W. C. (1974). The ratings of individuals in organizations: An alternative approach. Organizational Behavior and Human Performance, 12, 105-124.
- Cambell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.
- Carroll, S. J., & Schneier, C. E. (1982). Performance appraisal and review systems: The identification, measurement, and development of performance in organizations. Glenview, IL: Scott, Foresman.

- Cascio, W. F. (1991). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Clement, R. W., & Stevens, G. E. (1989). Performance appraisal in higher education: Comparing departments of management with other business units. Public Personnel Management, 18, 263-278.
- Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. Journal of Applied Psychology, 74, 130-135.
- Dornstein, M. (1989). The fairness judgments of received pay and their determinants. Journal of Occupational Psychology, 62, 287-299.
- Farh, J., & Dobbins, G. H. (1989). Effects of comparative performance information on the accuracy of self-ratings and agreement between self- and supervisory ratings. Journal of Applied Psychology, 74, 606-610.
- Farh, J., & 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.
- Farh, J., Werbel, J. D., & Bedeian, A. G. (1988). An empirical investigation of self-appraised based performance evaluation. Personnel Psychology, 41, 141-156.
- Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.

Fisher, C. D. (1989). Self and superior assessment: Unraveling the causes of disagreement. Unpublished manuscript. University of Baltimore, Baltimore, MD.

Fletcher, C. (1986). The effects of performance review in appraisal: Evidence and implications. Journal of Management Development, 5, 3-12.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. Personnel Psychology, 41, 581-592.

Goodman, P. S. (1974). An examination of the referents used in the evaluation of pay. Organizational Behavior and Human Performance, 12, 170-195.

Guba, E. G., & Lincoln, Y. S. (1989). Fourth Generation Evaluation. Newbury Park, CA: Sage.

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

Henderson, R. I. (1984). Performance appraisal. Reston, VA: Reston Publishing Company.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. Personnel Psychology, 39, 811-826.

Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. Personnel Psychology, 44, 601-618.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process in the 1980s: What has it contributed to appraisals in use? Organizational Behavior and Human Decision Processes, 54, 321-368.

Kulik, C. T., & Ambrose, M. L. (1992). Personal and situational determinants of referent choice. Academy of Management Review, 17, 212-237.

Landy, F. J. (1989). Psychology of work behavior. Pacific Grove, CA: Brooks/Cole.

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

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

Locher, A. H., & Teel, K. S. (1988). Appraisal trends. Personnel Psychology, 67, 139.

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

Magner, D. K. (1996). The faculty: Fewer professors believe western culture should be the cornerstone of the college curriculum. The Chronicle of Higher Education, 43(3), A12-A15.

McEnery, J., & McEnery, J. M. (1987). Self-ratings in management training needs assessment: A neglected opportunity? Journal of Occupational Psychology, 60, 49-60.

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

Murphy, K. R., & Cleveland, J. N. (1991). Performance appraisal: An organizational perspective. Boston: Allyn & Bacon.

Nathan, B. R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. Journal of Applied Psychology, 75, 290-296.

Oldham, G. R., Kulik, C. T., Ambrose, M. L., Stepina, L. P., & Brand, J. F. (1986). Relations between job facet comparisons and employee relations. Organizational Behavior and Human Decisions Processes, 38, 28-47.

Riggio, R. E., & Cole, E. J. (1992). Agreement between subordinate and superior ratings of supervisory performance and effects on self and subordinate job satisfaction. Journal of Occupational and Organizational Psychology, 65, 151-158.

Rosenthal, J. T., Cogan, M. L., Marshall, R., Meiland, J. W., Wion, P. K., & Molotsky, I. F. (1994). The work of faculty: Expectations, priorities, and rewards. Academe, 80, 35-48.

Schrader, B. W., & Steiner, D. D. (1996). Common comparison standards: An approach to improving agreement between self and supervisory performance ratings. Journal of Applied Psychology, 81, 813-820.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessments. Psychological Bulletin, 90, 322-351.

Somers, M. J., & Birnbaum, D. (1991). Assessing self-appraisal of job performance as an evaluation device: Are the poor results a function of method or methodology? Human Relations, 44, 1081-1091.

Sumers, T. P., & DeNisi, A. S. (1990). In search of Adams' other: Reexamination of referents used in the evaluation of pay. Human Relations, 43, 497-511.

Thorndike, R. L., (1949). Personnel selection: Test and measurement techniques. New York: Wiley.

Thornton, G. C. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-271.

Vecchio, R. P., & Gobdel, B. C. (1984). The vertical dyad linkage model of leadership: Problems and prospects. Organizational Behavior and Human Performance, 34, 5-20.

APPENDIX A
APPROVAL LETTER FROM
INSTITUTIONAL REVIEW BOARD



EMPORIA STATE UNIVERSITY

1200 COMMERCIAL EMPORIA, KANSAS 66801-5087 316/341-5351
Fax 316/341-5909

RESEARCH AND GRANTS CENTER - BOX 4003

March 21, 1997

Jamie Keiffer
909 Valley Circle
Emporia, KS 66801

Dear Ms. Keiffer:

The Institutional Review Board for Treatment of Human Subjects has evaluated your application for approval of human subject research entitled, "An Examination of the Interrater Agreement Between Self and Supervisory Performance Ratings in a Subject Occupation." The review board approved your application which will allow you to begin your research with subjects as outlined in your application materials.

Best of luck in your proposed research project. If the review board can help you in any other way, don't hesitate to contact us.

Sincerely,

A handwritten signature in cursive script, appearing to read "John O. Schwenn".

John O. Schwenn, Dean
Graduate Studies and Research

pf

cc: Brian Schrader

APPENDIX B
TRANSMITTAL LETTER

FOR FACULTY MEMBERS

Dear Faculty Member,

I am a graduate student at Emporia State University in Emporia, Kansas. I am working towards my master's degree in Industrial/Organizational Psychology. As partial fulfillment of my degree requirements, I am conducting a thesis project for which I am requesting your participation.

This study is intended to examine the basis on which you evaluate your work performance. Specifically, I am interested in increasing the agreement between self- and supervisory ratings in the performance appraisal process. If you are willing to participate, your responses will be kept completely confidential. The purpose of listing you and your chairperson's name is only to ensure that faculty members and chairpersons can be matched together for research purposes. Therefore, please be as honest as possible in your response since I am interested in clearly understanding the performance appraisal process in a university setting. The entire questionnaire should take approximately 5 minutes to complete.

I would greatly appreciate it if you would complete the packet and return it in the enclosed stamped, self-addressed envelope by **May 8, 1997**. I realize your schedule is busy and your time is valuable, but your responses will help further explain the process of evaluating employee work performance.

If you have any questions or comments regarding this research, you may contact me at (316) 341-5803 or (316) 343-1187. I want to thank you in advance for your support and cooperation.

Sincerely,

Jamie Kieffer
Graduate Student

FOR CHAIRPERSONS

Dear Chairperson,

I am a graduate student at Emporia State University in Emporia, Kansas. I am working towards my master's degree in Industrial/Organizational Psychology. As partial fulfillment of my degree requirements, I am conducting a research project for which I am requesting your participation.

This study is intended to examine the basis on which you evaluate employee performance. Specifically, I am interested in increasing the agreement between self- and supervisory ratings in the performance appraisal process. If you are willing to participate, your responses will be kept completely confidential. The purpose of listing you and your faculty member's name is only to ensure that faculty members and chairpersons can be matched together for research purposes. Therefore, please be as honest as possible in your response since I am interested in clearly understanding the performance appraisal process in a university setting. The entire questionnaire should take approximately 5 minutes to complete.

After receiving this survey, please select one faculty member who has tenure, is working in your department, and has been in his/her current position for at least two years. Please give the other part of this packet to that faculty member so that he/she can complete the survey. The section which has the letter addressed to "faculty member" is the part you will distribute to the tenured employee. Note that the faculty member you select will also be the individual you evaluate when answering your part of the questionnaire.

I would greatly appreciate it if you would complete your part of the packet and return it in the enclosed stamped, self-addressed envelope by **May 8, 1997**. I realize your schedule is busy and your time is valuable, but your responses will help further explain the process of evaluating employee work performance.

If you have any questions or comments regarding this research, you may contact me at (316) 341-5803 or (316) 343-1187. I want to thank you in advance for your support and cooperation.

Sincerely,

Jamie Kieffer
Graduate Student

APPENDIX C
PACKET INSTRUCTIONS AND
INFORMED CONSENT DOCUMENT

FOR FACULTY MEMBERS

Your Name_____
Chairperson's NamePERFORMANCE APPRAISAL QUESTIONS
AND RATINGS PACKET

The Division of Psychology and Special Education of Emporia State University supports the practice of protection for human subjects participating in research and related activities. The following information is provided so that you can decide whether you would like to participate in the present study.

You are going to be asked to fill out a series of questions and rating scales pertaining to your job performance and on what basis you evaluate your performance. It is VERY IMPORTANT that you DO NOT look ahead; proceed one page at a time. Please provide honest and accurate responses for all questions and ratings. The entire packet should take approximately 5 minutes to complete.

All responses made in this packet will remain confidential and will be used for research purposes ONLY. Your individual responses WILL NOT be made available to your chairperson, your co-workers, or your university. The purpose of listing you and your chairperson's name at the top of this sheet is only to ensure that faculty members and chairpersons can be matched together for research purposes.

By signing and dating this form, you are providing your voluntary consent to participate in this research (by completing the remainder of the packet) as described above.

I, _____ have read the above information and
(please print name)
have decided to participate. I understand that my participation is voluntary.

Signature_____
Date

FOR CHAIRPERSONS

Your Name_____
Faculty Member's NamePERFORMANCE APPRAISAL QUESTIONS
AND RATINGS PACKET

The Division of Psychology and Special Education of Emporia State University supports the practice of protection for human subjects participating in research and related activities. The following information is provided so that you can decide whether you would like to participate in the present study.

You are going to be asked to fill out a series of questions and rating scales pertaining to your job performance and on what basis you evaluate your performance. It is VERY IMPORTANT that you DO NOT look ahead; proceed one page at a time. Please provide honest and accurate responses for all questions and ratings. The entire packet should take approximately 5 minutes to complete.

All responses made in this packet will remain confidential and will be used for research purposes ONLY. Your individual responses WILL NOT be made available to your faculty member or your university. The purpose of listing you and your faculty member's name at the top of this sheet is only to ensure that faculty members and chairpersons can be matched together for research purposes.

By signing and dating this form, you are providing your voluntary consent to participate in this research (by completing the remainder of the packet) as described above.

I, _____ have read the above information and have
(please print name)
decided to participate. I understand that my participation is voluntary.

Signature_____
Date

APPENDIX D
RATING INSTRUCTIONS AND PERFORMANCE DIMENSIONS

RATING SHEET INSTRUCTIONS

The next six pages will be asking you to make ratings across three performance dimensions. The six rating sheets are exactly identical EXCEPT for the instructions on how to generate your ratings. It is VERY IMPORTANT that you read the instructions at the top of each page carefully and provide ratings in a manner consistent with the specific instructions. Listed below are the definitions of what constitutes a **very specific** facet of the three dimensions. The researcher recognizes that these are not all-inclusive definitions and only cover some the aspects of the larger dimension. However, they meet the research needs of the current study.

- DIMENSION 1:** **Scholarship** - Publications printed in a professional journal, book, or chapter
- DIMENSION 2:** **Service** - Participation and involvement in distinct university or department committees at your university
- DIMENSION 3:** **Teaching/Instruction** - Student evaluations of your teaching methods/techniques

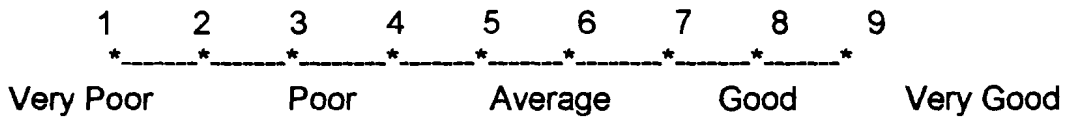
APPENDIX E
FACULTY MEMBER RATING SHEETS

AMBIGUOUS

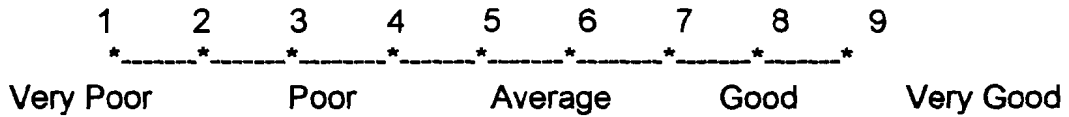
Based on your performance over the past two years, please rate yourself on the following performance dimensions.

Please circle the appropriate number for each dimension.

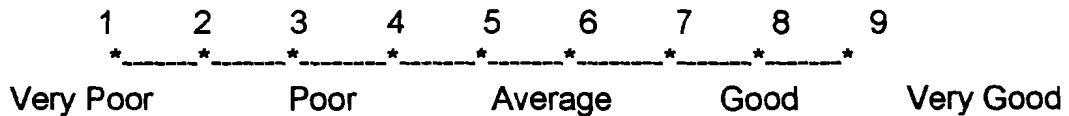
SCHOLARSHIP



SERVICE



TEACHING/INSTRUCTION



INTERNAL

Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use your own personal, internal values and standards as a criteria. That is, base your ratings on how well you personally feel you have done in the past year relative to your abilities and past performance. DO NOT give consideration to any other criteria beyond you **own beliefs** as to how well you performed.

Please circle the appropriate number for each dimension.

SCHOLARSHIP

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

SERVICE

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

TEACHING/INSTRUCTION

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

ABSOLUTE

Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use the average requirement or goal listed in the parentheses next to each dimension as a criteria. While different universities have different expectations of their employees, please make your rating against the defined average whether appropriate for your department/institution or not. That is, for each dimension rate yourself in comparison to the average level of performance of tenured faculty members. **DO NOT** give consideration to any other criteria beyond your own belief as to whether or not you met this requirement.

Please circle the appropriate number for each dimension. The averages listed below were contrived for the purposes of this study and do not necessarily indicate appropriate averages for all universities, but please base ratings on the averages listed below.

SCHOLARSHIP (The average number of publications for a tenured faculty member over a two year period is 3)

| | | | | | | | | |
|-----------|------|---|---------|---|---|------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Poor | Poor | | Average | | | Good | | Very Good |

SERVICE (The average number of different committees a tenured faculty member serves over two years is 4)

| | | | | | | | | |
|-----------|------|---|---------|---|---|------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Poor | Poor | | Average | | | Good | | Very Good |

TEACHING/INSTRUCTION (The average student evaluation rating of a tenured faculty member is 4.0 on a 5-point scale)

| | | | | | | | | |
|-----------|------|---|---------|---|---|------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Poor | Poor | | Average | | | Good | | Very Good |

RELATIVE - INSIDE

Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use the performance of fellow faculty members who work within your university as a criteria. That is, think about how other university faculty have performed and compare yourself to them. **DO NOT** give consideration to any other criteria beyond your own belief as to how well you performed in **direct comparison to fellow faculty members within your university.**

Please circle the appropriate number for each dimension.

SCHOLARSHIP

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

SERVICE

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

TEACHING/INSTRUCTION

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

RELATIVE - OUTSIDE

Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use the performance of other individuals who have similar jobs and who work in comparable departments, but work outside your university as a criteria. That is, think about how others who have similar jobs have performed and compare yourself to them. **DO NOT** give consideration to any other criteria beyond your own belief as to how well you performed **in direct comparison to other individuals with similar jobs who work outside your university.**

Please circle the appropriate number for each dimension.

SCHOLARSHIP

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

SERVICE

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

TEACHING/INSTRUCTION

| | | | | | | | | |
|-----------|------|---------|---------|---------|---|---------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | *-----* | | *-----* | | *-----* | | * |
| Very Poor | Poor | | Average | | | Good | | Very Good |

MULTIPLE

Based on your performance over the past two years, please rate yourself on the following performance dimensions. Use your own personal standards, your attainment of the average requirements and goals, and your comparison with other faculty members both within and outside your university as the criteria. That is, consider all four standards as defined in the previous pages. Give equal consideration to all four of the criteria.

Please circle the appropriate number for each dimension.

SCHOLARSHIP

| | | | | | | | | |
|-----------|------|---|---------|---|---|------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Poor | Poor | | Average | | | Good | | Very Good |

SERVICE

| | | | | | | | | |
|-----------|------|---|---------|---|---|------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Poor | Poor | | Average | | | Good | | Very Good |

TEACHING/INSTRUCTION

| | | | | | | | | |
|-----------|------|---|---------|---|---|------|---|-----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Poor | Poor | | Average | | | Good | | Very Good |

APPENDIX F
CHAIRPERSON RATING SHEETS

AMBIGUOUS

Based on your faculty member's performance over the past two years, please rate this employee on the following performance dimensions.

INTERNAL

Based on your faculty member's performance over the past two years, please rate this employee on the following performance dimensions. Use your perceptions of the faculty member's own personal, internal values and standards as the criteria. That is, base your ratings on how you think the faculty member feels they have done over the past year relative to their abilities and past performance. **DO NOT** give consideration to any other criteria beyond how **you believe the employee perceives they have done** over the past year.

ABSOLUTE

Based on your faculty member's performance over the past two years, please rate this employee on the following performance dimensions. Use the average requirement or goal listed in the parentheses next to each dimension as a criteria. While different universities have different expectations of their employees, please make your rating against the defined average. That is, for each dimension rate the faculty member in comparison the average level of performance as defined below. **DO NOT** give consideration to any other criteria beyond your own belief as to **whether or not the employee met this requirement**.

(appendix continues)

(APPENDIX F continued)

RELATIVE - INSIDE

Based on your faculty member's performance over the past two years, please rate this employee on the following performance dimensions. Use the faculty member's fellow faculty members who work within your university as the criteria. That is, think about how other university faculty members have performed and compare the faculty member to them. **DO NOT** give consideration to any other criteria beyond your own belief as to how well the faculty member performed **in direct comparison to his/her fellow faculty members within your university.**

RELATIVE - OUTSIDE

Based on your faculty member's performance over the past two years, please rate this employee on the following performance dimensions. Use the performance of other individuals who have similar jobs and who work in comparable departments, but work outside your university as a criteria. That is, think about how other individuals outside your university who have similar jobs have performed and compare the faculty member to them. **DO NOT** give consideration to any other criteria beyond your own belief as to how well the faculty member performed **in direct comparison to other individuals with similar jobs who work outside your university.**

MULTIPLE

Based on your faculty member's performance over the past two years, please rate this faculty member on the following performance dimensions. Use

(appendix continues)

(APPENDIX F continued)

your perceptions of the employee's own personal standards, the faculty member's attainment of the average requirements and goals, and the comparison with other faculty members both inside and outside your university as the criteria. That is, consider all four standards as defined in the previous pages. Give equal consideration to all four of the criteria.

APPENDIX G
POST-RATING COMPARISON STANDARD QUESTIONS

FOR CHAIRPERSONS

Please think about how you have rated your faculty member's job performance prior to answering this packet. Based on the previous five comparison standards (**internal**, **absolute**, **relative-inside**, **relative-outside**, and **multiple**), which one have you used most often in the past as the basis for your ratings. That is, which one have you used to decide whether or not the faculty member was performing satisfactorily on the job? You may refer back to the comparison standard instructions on the previous rating sheets if you need to.

Please circle the comparison standard that you have previously used.

INTERNAL STANDARD (Own internal values and standards)

ABSOLUTE STANDARD (Average performance of university faculty member)

RELATIVE-INSIDE (Performance of co-workers working inside your university)

RELATIVE-OUTSIDE (Performance of individuals with similar jobs working outside your university)

MULTIPLE STANDARD (Combination of previous standards)

FOR CHAIRPERSONS

If asked to rate a faculty member's performance in the future, please rate each of the five comparison standards as your preference for using them in future performance appraisals. You may refer back to the comparison standard instructions on the previous rating sheets if you need to. *Please circle the appropriate number for each standard.*

INTERNAL STANDARD (Own internal values and standards)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

ABSOLUTE STANDARD (Average performance of university faculty member)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

RELATIVE-INSIDE (Performance of co-workers working inside your university)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

RELATIVE-OUTSIDE (Performance of individuals with similar jobs working outside your university)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

MULTIPLE STANDARD (Combination of previous standards)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

FOR FACULTY MEMBERS

If asked to evaluate your own job performance in the future (i.e., provide a self-rating), please rate each of the five comparison standards as to your preference for using them in future performance ratings. You may refer back to the comparison standard instructions on the previous rating sheets if you need to. *Please circle the appropriate number for each standard.*

INTERNAL STANDARD (Own internal values and standards)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

ABSOLUTE STANDARD (Average performance of university faculty member)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

RELATIVE-INSIDE (Performance of co-workers working inside your university)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

RELATIVE-OUTSIDE (Performance of individuals with similar jobs working outside your university)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

MULTIPLE STANDARD (Combination of previous standards)

| | | | | | | | | |
|---------------------|---|----------------|---|---------|---|-----------------|---|----------------------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *-----* | | | | | | | | |
| Very Low Preference | | Low Preference | | Neutral | | High Preference | | Very High Preference |

APPENDIX H
DEMOGRAPHICS AND COMPREHENSION QUESTION

DEMOGRAPHIC PROFILE

Age: _____

Sex: _____

Current Job Title/Occupation: _____

Number of years with your university: _____

Do you have tenure at your university? _____

(FOR CHAIRPERSONS ONLY)

Number of faculty members under your supervision: _____

Do you feel you understood all the instructions and questions asked throughout this packet and were able to answer them in an honest and accurate manner?

YES or NO

I, Jamie C. Kieffer, hereby submit this thesis to Emporia State University as partial fulfillment of the requirements for an advanced degree. I agree that the Library of the University may make it available for use in accordance with its regulations governing materials of this type. I further agree that quoting, photocopying, or other reproduction of this document is allowed for private study, scholarship (including teaching) and research purposes of a nonprofit nature. No copying which involves the potential financial gain will be allowed without the written permission of the author.

Jamie C. Kieffer
Signature of Author

July 27, 1997
Date

An Examination of the Interrater Agreement
Between Self- and Supervisory Performance
Ratings in a Subjective Occupation

Title of Thesis

Day Cooper
Signature of Graduate Office Staff Member

July 27, 1997
Date Received