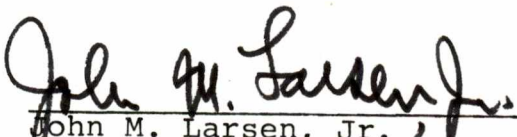
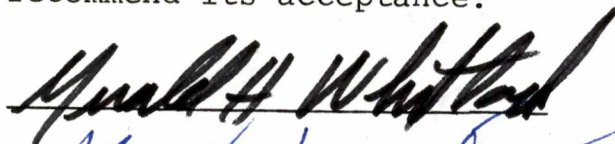
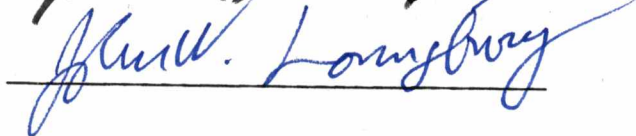


To the Graduate Council:

I am submitting herewith a thesis written by Marvin Lee Stewart entitled "Reliability of a Performance Evaluation Process and the Effect of Confidence Ratings on Reliability." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Industrial and Organizational Psychology.


John M. Larsen, Jr.,
Major Professor

We have read this thesis and
recommend its acceptance:

Accepted for the Council:


Vice Chancellor
Graduate Studies and Research

RELIABILITY OF A PERFORMANCE EVALUATION PROCESS AND
THE EFFECT OF CONFIDENCE RATINGS ON RELIABILITY

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Marvin Lee Stewart

June 1982

3060950

To Geneva

ACKNOWLEDGMENTS

My gratitude is extended to Dr. J. W. Lounsbury and Dr. G. H. Whitlock for serving on my committee. A special note of thanks is given to Dr. J. M. Larsen, Jr., for chairing my committee and providing the encouragement and guidance in this project. Without his interest and support this thesis would have not become reality. I would also like to thank my wife, Geneva, for her patience, understanding and encouragement. She never lost faith in my ability to complete this project and was willing to make many personal sacrifices to help me with the writing of this thesis. Also a special thanks to Alexander, my son, who brought a new enthusiasm into my life. I would like to thank my parents for their support in all of my endeavors, especially my educational ones.

I would like to extend my appreciation and gratitude to Bill Redding and his staff for allowing me to enter their organization and collect the data for this thesis. Mr. Redding and his staff were always ready and willing to assist me in any way they could.

ABSTRACT

The purpose of this study was to examine the reliability of a performance evaluation form in an on-the-job setting. Also this study was used to determine if the use of a confidence rating scale, in addition to the performance rating scale, could be used as an indicator of the opportunity of the rater to observe performance. A 54 item rating form using the summated rating format was developed through a series of interactions between a professional consultant and the users of the form in a manner similar to the retranslation method of scale development. The rating form was used to rate the performance of 143 job incumbents in the circulation department of a newspaper. Two sets of ratings were obtained, one by the first line supervisor and a second by the department head.

Interrater reliability coefficients were computed for items separately, for subgroups of items, for "Overall" subgroup items taken together, and for the total overall form. In all but a few exceptions the reliability coefficients were higher than expected in the light of other field studies, especially those studies using two different levels of management as in this study. The total overall interrater reliability coefficient for the performance rating was .65. Alpha coefficients were also computed for the

subgroups of items and total overall form for two levels of raters. These were very high indicating a high internal consistency for the rating form.

To examine the use of the confidence rating the interrater reliability coefficients for items with high rater confidence scores were compared with the reliability coefficients for items with no restriction on the confidence rating. This was also repeated where the interrater reliability coefficients for items with a low department head confidence rating were used instead of a high confidence rating. In each case confidence rating had no relationship to higher or lower reliability coefficients respectively.

The high interrater reliability coefficients in this study would indicate that a relatively high degree of reliability can be obtained in an on-the-job setting if reasonable care is taken with the development of a performance rating form and in training and instructing the raters who use it. The results of this study indicated that the confidence ratings had no apparent relationship to reliability indices in this case and are of little value.

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	1
The Problem	2
Statement of the problem	2
The importance of the study	3
Scope of the study	5
II. LITERATURE REVIEW	11
Introduction	11
Rater Errors	13
Rating Format Problems	19
Rater Performance Observation Problems	21
Summary	22
III. METHODS AND PROCEDURES	25
Sample	25
Instrument	26
Data Collection	28
Analysis	30
IV. RESULTS	33
V. DISCUSSION OF RESULTS	53
VI. SUMMARY, CONCLUSION, AND IMPLICATIONS	63
Summary	63
Conclusions	67
Implications	70
BIBLIOGRAPHY	72
APPENDIXES	78
Appendix A	79
Appendix B	92
Appendix C	101
VITA	108

LIST OF TABLES

TABLE	PAGE
1. Performance Item Correlations of First Line Supervisor Ratings with Department Head Ratings	34
2. Predicted Performance Item Correlations of First Line Supervisor Ratings with Department Head Ratings	36
3. Performance Item Subgroup Correlations of First Line Supervisor Ratings with Department Head Ratings	40
4. Predicted Performance Item Subgroup Correlations of First Line Supervisor Ratings with Department Head Ratings	41
5. Comparison of Item Correlations Having High Department Head Confidence Rating with Item Correlations Having Unrestricted Confidence Rating	43
6. Comparison of Item Correlations Having Low Department Head Confidence Rating with Item Correlations Having Unrestricted Confidence Rating	46
7. Comparison of Item Correlations Having High Department Head Confidence Rating with Item Correlations Having Low Department Head Confidence Rating	48
8. Comparison of Item Correlations Having High Department Head and Supervisor Confidence Ratings with Item Correlations Having Unrestricted Confidence Ratings	50
9. Item Means and Standard Deviations for Department Head Performance Ratings	93
10. Item Means and Standard Deviations for First Line Supervisor Performance Ratings	95
11. Item Means and Standard Deviations for Department Head Predicted Performance Ratings	97

TABLE

PAGE

12.	Item Means and Standard Deviations for First Line Supervisor Predicted Performance Ratings	99
13.	Item Correlations for Performance Ratings with Confidence Ratings for First Line Supervisor . .	102
14.	Item Correlations for Performance Ratings with Confidence Ratings for Department Head	104
15.	Item Correlations for Performance Ratings with Confidence Ratings for Both Supervisors Combined	106

CHAPTER I

INTRODUCTION

An important function of management in an organization is to judge the performance of the organization and the people within the organization. The evaluation of the individual's performance will become the basis for such decisions as promotion, employee selection, wages, conformance to Equal Employment Opportunity Commission (EEOC) guidelines, and many other personnel decisions. The most common method of individual job performance evaluation has been to have another peer or supervisor rate that individual's performance. Because this is the most common method, much of the research on job performance evaluation has centered on job performance rating methods and the problems associated with these techniques.

Most of the problems in job performance ratings are associated with validity and reliability of the rating methods. Borman (1978) summarized many of the questions and the research approaches to the problems of job performance ratings. Rater errors and a lack of discriminant validity were indicated as the most frequent obstacles in developing better performance ratings. Rater errors and a lack of discriminant validity were indicated as the most frequent obstacles in developing better performance ratings. Borman

went on to point out four general categories of causes for these performance rating problems. They were: (1) Does the rater have ample opportunity to observe relevant job behavior of the job dimension being rated? (2) Is the rater familiar with the types of rater error and does the rater have the training to know how to avoid them? (3) Does the rating format provide a means to translate observed behavior into measurable responses? (4) Is the organizational climate such that the rater can make an unbiased judgement free of pressure (real or imaginary)? Borman examined these questions in the laboratory setting where most of the variables could be controlled and manipulated. This was an attempt to set the "upper limits of validity and reliability" for performance ratings. Borman's research was from a scientist's theoretical approach. The present research was an examination of might be expected in terms of reliability in a field study from a practitioner's approach.

I. THE PROBLEM

Statement of the Problem

The purpose in this study was to examine the reliability of a performance rating by comparing the ratings of a first line supervisor and a department head for 123 job incumbents in an on-the-job setting. Also included in the study were the effects on reliability of high and low rater confidence

ratings for the items on the performance rating questionnaire. As part of the examination of the reliability of the questionnaire, internal consistency of the questionnaire was evaluated in order to determine the possible extent of "halo" rater error.

The emphasis throughout this study was that it was done in the field with on-the-job personnel and that the measurement characteristics of the instrument and the process of job performance evaluation were important.

The Importance of the Study

As mentioned above, performance ratings are essential in making many management and personnel decisions. Often the performance rating is the only information available regarding performance whether it is valid and reliable or not. The industrial and organizational psychologist would like to have performance ratings done under highly controlled conditions such as Borman used in his research where reliability, validity and rater errors can be checked and controlled. The field situation is more likely to involve poor rater training, inadequate checks on reliability, inadequate opportunities to observe performance, and inappropriate use of the rating format. Thorndike and Hagen (1961) indicated in their research that accurate (reliable) ratings depend upon rater's opportunity to observe relevant job performance behavior. Lacking opportunity to observe,

of course, can leave the decision maker with something worse than no information and information that is possibly false or misleading. If it is not possible to set an ideal rating environment, an option would be to determine the measurement characteristics of the rating based upon the opportunity to observe behavior. The present research was aimed at examining such a solution using confidence scores as such an indicator of opportunity to observe. If this index of confidence could be used in a field setting, using trained managers and supervisors and using a reliable form one could have some confidence in the measure as actually measuring performance behavior.

This research also provided another opportunity to observe what Bernardin (1977) would have called a "summated rating scale" in use in a field setting (a 7 point Likert type scale in this case). The summated rating score format tends to result in increased reliability because the rater can respond in more than just a yes or no mode (Borman, 1979). In this research it was possible to examine the reliability of a job performance evaluation where such a format was used for two types of ratings (actual job performance and predicted job performance) by raters from two different levels of management. Thus just as sample size can be increased to decrease importance of limiting assumptions in a statistical test, multiple testing of a rating format in different

situations decreases limiting assumptions and could increase confidence in the reliability of the ratings. The more a particular rating format is used in research such as this, with reliable results, the more likely there will be increased confidence in that rating format.

In general, this research was useful because it added information for field tested performance ratings. Here is a test of the reliability of performance ratings under actual conditions in a working organization where ratings had to be made under the influence of many unknown factors operating on the rater. Factors such as pressure from superiors, less than comprehensive training in avoidance of typical rater errors, variability in the opportunity to observe behavior, time differentials in making ratings, and confusion regarding rating items that may pertain only marginally to the job are important. Here then was an opportunity to assess the reliability of this performance rating format under actual working conditions.

Scope of the Study

In 1980 a newspaper publishing company in a Southeastern city was seeking a method of evaluating the job performance of distribution personnel in the circulation department. These are the people who supervise the local paper carriers, distribution, and sales and provide the communications link

with customers for the newspaper. The personnel department sought the aid of a consultant to develop the procedures and instruments. It was anticipated that if a reliable instrument and process could be developed, results would serve to assist management in personnel decisions and in the development of a selection program.

It was decided that the best way to approach the performance evaluation process was to have two levels of immediate superiors rate the subordinates using a rating questionnaire with summated scale format. The items for this questionnaire were to be based on specific behavioral job activities that were considered important to the job by management and employees. To this end, extensive interviews were conducted with supervisors and managers, job descriptions were examined, the consultant travelled on work days with district personnel, and old documents were reviewed. A questionnaire was constructed and interviews were undertaken to revise the form. A revised draft form was used to meet with all supervisors to use in training. However, the intent was to have the supervisors revise, review, reword and generally work through the form. A second draft revision was similarly reworked and a third and fourth as well. When the supervisors and managers were satisfied with the form, training on the use of the form continued.

The training included a full outline of known possible difficulties, biases, errors, procedures, etc. and practice

ratings from case descriptions. The supervisors thus were well trained in job performance evaluations, but they had not completed a performance evaluation form for real use until this time.

The first level supervisors, after the training in the form's use, used the final form to evaluate people, essentially in three job titles: District Manager, Relief District Manager, and Utility District Manager as well as to appraise persons in other similar or related jobs dealing with single copy circulation such as news rack manager. A total of about 143 people was evaluated.

To provide the necessary second rater for test of the questionnaire's reliability the department heads were asked to perform the same ratings using the questionnaire. When the two ratings were paired it was found that the department heads did not rate about 20 of the original 143 ratees because they worked in remote locations and had only indirect contact with them through the first line supervisors. These remaining 123 pairs of performance ratings provided the basic data for analysis in this study.

The questionnaire used for the job performance ratings consisted of 54 items divided into subgroups that represented the major job dimensions of: service, sales, accounting and record keeping, and effort. Each of these groups was divided into smaller groups of items representing more specific

aspects of the major dimensions. Each ratee was rated on each item on a scale from 1 (superior) to 7 (unacceptable). This scale was given operational meaning with a definition of each scale point provided on the cover page set of instructions (see Appendix A). Also associated with each of these 54 items was a scale to indicate the confidence the rater had in the rating given. This was a four point scale with points ranging from A (very high confidence) to D (no confidence). These scale points were also defined on the cover page instructions (see Appendix A).

The second part of the questionnaire was a 44 item section that asked the rater to predict future performance in terms of speed in learning the job. A seven point scale similar to the performance rating was used for these ratings.

One of the basic purposes in this research was to estimate the reliability of the job performance evaluations in the actual job setting. The performance rating scores given by the ratee's first line supervisor were correlated with the scores given by the department head. This reliability coefficient, an estimate of interrater reliability, was then compared with the estimated reliability reported in the literature regarding other performance evaluation processes.

Another aspect in this research was to utilize the confidence ratings incorporated in the form. Correlations

of the ratings made by the two raters for each ratee were compared in two ways. Interrater agreement indices, item by item, major dimension by major dimension and overall were computed. In addition the same item by item index calculations were performed using only those items for which the department head had indicated high (A or B) confidence rating and using only those items for which the department head had indicated low (C or D) confidence rating. A Fisher Z transformation was used to make the comparisons between reliability indices with a high confidence rating and those with low confidence ratings.

The internal consistency of the questionnaire was examined by computing Cronbach's alpha coefficient (Nunnally, 1967) for the questionnaire as a whole, for the group sub-totals and for each item.

There are two important research questions to be addressed in the present research. One question is what is the reliability of the job performance evaluation form as used in this study. The second question is in the form of a set of hypotheses. It was hypothesized that the interrater reliability coefficients for items with high department head confidence ratings would be higher than the interrater reliability coefficients for items with unrestricted confidence ratings. Another hypothesis was that the interrater reliability coefficients for items with low department

head confidence ratings would be lower than the interrater reliability coefficients for items with unrestricted confidence ratings. A third hypothesis was that the interrater reliability coefficients for items with high department head confidence ratings would be higher than the interrater reliability coefficients for items with low department head confidence ratings. A final hypothesis was that the interrater reliability coefficients for items with high department head and supervisor confidence ratings would be higher than the interrater reliability coefficients for the items with no restriction put on the confidence ratings. The purpose of these four hypotheses was to determine if the rater's opportunity to observe job performance behavior could be indicated on the confidence rating scale and if differences in the interrater reliability coefficients would occur.

CHAPTER II

LITERATURE REVIEW

I. INTRODUCTION

Sir Francis Galton and his colleague Karl Pearson did some of the earliest work in devising "methods for analyzing the correlations between different measurement methods," (Nunnally, 1970). But it was Charles Spearman and his associates who laid down many of the principles that are still the basis for much of the theory of measurement error or reliability theory. [Nunnally (1967) defines reliability as the "extent to which measurements are repeatable—by the same individual using different measures of the same attribute or by different persons using the same measure of an attribute."] Spearman made the measurement of reliability so clear that often the concern for reliability overshadows many other important concerns, such as validity, in psychological research. On the other hand, reliability should not be disregarded but should share equally with other considerations in a research or measurement design according to Nunnally (1967).

Once the measurability and importance of measurement error had been established many researchers began to investigate the causes of the unreliability in measurement and

ways to avoid these errors. Barrett (1966) outlined what were some of the problems and causes of unreliability in performance measures or in ratings used as criteria in research. They were mostly in the form of rater errors such as halo, leniency/strictness, and central tendency. Barrett also noted that seeking a highly reliable measure was unjustified if the result was a trivial measure that showed no validity with the criteria. Another point discussed was that pressure on a rater to be consistent may force some raters to refer surreptitiously to previous ratings. Borman (1978) organized his problems and causes of unreliability in performance measures into four categories. First were the problems with reliability that were associated with raters not having observed the behavior they are supposed to rate. Another area of concern, like Barrett's, was rater errors such as halo, leniency, and central tendency and how to avoid them. Borman's third grouping was under a heading called rating format problems. These concerned what form a written rating questionnaire took. The last area in Borman's organization of causes of reliability problems was labeled organizational constraints. This concerned the effects of such organizational groups as unions, peers, and management on reliability problems.

Many researchers have organized problems with reliability in other ways but most may be included in headings mentioned

above. Researchers most often choose one problem with reliability and seek to explain it and find ways of reducing it. In the next part of this literature review is reported some of the research that has been done with these specific problems and causes of unreliability in performance measures.

II. RATER ERRORS

One reliability problem associated with rater error is halo. Thorndike (1920) was one of the first researchers to address the problem and define it by saying

. . . ratings are apparently affected by a marked tendency to think of a person in general as rather good or rather inferior and to color judgments of qualities by this general feeling.

Barrett (1966) discusses halo and finds that trying to interpret halo may be a mistake because often traits are positively related. For more support Bingham (1939) argues that there is valid halo in some cases because the rater may see an overall quality of good performance that is a combination of positive and negative individual qualities. Asch (1946) found support for Bingham by finding that forming an impression is an organized process and depends on a dynamic relationship of several individual scales. (It should be noted that many of these early studies dealt with personality traits.) On the other hand, Cummings and Schwab (1973) finds that characteristics of individuals have no relation.

Cummings and Schwab say, "We would expect little relationship between various performance subcomponents of an individual being observed." Support for this position came from a book by Anastasi (1958) which is a summary of theories and research on psychological differences between people. "Good" or "bad," halo continues to be reported in the literature in such as Campbell, Dunnette, Lawler, and Weick (1970), Anderson, Roush, and McClary (1973), Warmke and Billings (1979), Latham, Wexley and Pursell (1975), Bernardin and Walter (1977), Bernardin (1978), Borman (1978), Borman (1979). From the above research it is unclear how much "halo" is acceptable and how much is unacceptable but most researchers agree that if dimension intercorrelations are too high then steps need to be taken to lower it.

Barrett (1966) suggests that halo may be reduced where several ratees are being rated at the same time by rating all ratees on one dimension of performance before going on to another dimension. He offers no empirical support for this. Similar to this approach is the use of the mixed standard scale format questionnaire for rating performance (Blanz and Ghiselli, 1972) where the items on the questionnaire are mixed with no regard to performance dimensions. This type of scale has shown some promise in reducing halo rating errors. Most efforts in reducing halo have been in offering training to the raters to make them aware of halo rating

errors and suggestions to avoid halo errors (Brown, 1968; Taylor and Hastman, 1956; Borman, 1975).

The rater error of leniency as described by Barrett (1966) is usually where "the word average has come to be an insult." Bass (1956) and Evans (1950) found many reasons for leniency errors in rating. For one example Porter, Lawler, and Hackman (1975) and McGregor (1957) discuss "the incompatible roles of judge and counselor" for the raters who make pay and promotion judgments and still work with and counsel the ratees. Mandell (1956) suggested that individual personality contributes to the amount of leniency in a rater's ratings. He describes a lenient rater as older, liking people, uncritical, self-confident, and secure. Taylor and Wherry (1951) found that when ratings would effect a person's future the rater was more lenient than in an experimental setting. Stockford and Bissel (1949) demonstrated that raters were more lenient if they thought the ratees would see the rating. Borman (1978) places, "organizational constraints on raters to provide performance evaluations different from their best estimates of ratees' true performance," in a separate division from leniency errors but in general it might be just another reason for leniency since most of the time organizational demands are for higher ratings for some of the reasons mentioned above.

In an attempt to reduce leniency errors some researchers have looked to training programs for the raters while other

researchers have sought other solutions. Levine and Butler (1952) found a reduction in leniency for ratings where a training program was used. Bittner (1948) in his research using 603 officers rating 2401 men showed a reduction in leniency for trained raters when compared with untrained raters. In a study by Bernardin (1978) it was observed that raters receiving an hour training session on rater errors had less halo and leniency errors in their ratings than raters who received a very short session of training. It was also noted that the training effect dissipated over time.

While there has been much support for rater training to reduce rater errors, there are several studies indicating that rater training has less than expected effects on rater errors. Taylor and Hastman (1956) found no difference in trained and untrained raters in the quality of ratings. Barrett (1966) reports a study by Katzell where training actually caused a deterioration in the quality of the ratings. Along this same theme Crow (1957) and Borman (1975) suggest from their research that training raters about the common rater errors may increase other reliability and accuracy problems. For example asking a rater to "spread their ratings over the scale" and "reduce the number of ratings concentrated at the high or low end" may cause ratings not to reflect actual performance. Borman (1979) in a later study found

only "limited support for previous results showing that persons can be trained to reduce psychometric errors in their ratings." Still, even with these problems in training it is common opinion that training can aid in reducing rater errors.

One of the alternate approaches to reducing leniency errors in ratings has been to change the rating format to one of "forced-choice." Sisson (1948) reported a reduction in leniency when forced-choice ratings were used for Army ratings. In this study the expected relationship between rank and efficiency was less using the forced-choice rating than using the previous rating scale which was interpreted as a decrease in leniency error. Another rating format comparison was made by Bernardin, Alvares, and Cranny (1976). Leniency error was reduced by using a summated scale as compared to a behavioral expectation scale format.

A third approach to reduce leniency error was used by Bernardin and Walter (1977). They had one group of college students keep a diary on the teaching performance of instructors and then rate performance at the end of the term. These ratings had less leniency errors in them than a similar group that rated performance without keeping a diary.

Another rater error is termed "central tendency." This is a tendency for raters to place most of their ratings in the center of a scale. Barrett, Taylor, Parker and Martens

(1958) concluded from a group of studies with high peaks at the center of the rating distributions that the unclear instructions were the cause. Barrett (1966) relates an example told him by Hoberman where organizational pressures caused raters to concentrate their ratings in the center of the scale.

The causes of a central tendency rater error are the same as for leniency problems in some cases. It may be caused by ambiguity in scale definitions (Barrett et al., 1958). Another possible cause can be lack of motivation by the rater (Taft, 1955). If the rater sees no purpose in a rating he rates everyone either high or average for the sake of time and convenience. Often the central tendency error is approached in discussions of discriminant validity.

The efforts made to reduce central tendency errors are the same as those made seeking to reduce leniency errors. As mentioned above they are the training of raters, changing of the rating format, improving observation opportunity, and trying to reduce organizational pressures. The same problems also exist with using these methods to reduce central tendency errors. If, in the course of training raters, they are told about the normal distribution of scores, the raters may alter an otherwise accurate set of ratings to fit an expected set (Crow, 1957).

III. RATING FORMAT PROBLEMS

Much of the recent research into the problems of reliability has been in the area of developing a rating format that will increase reliability. Many rating methods have been developed such as graphic rating scales, ranking, paired comparison, forced distribution, forced choice, critical-incidents, weighted random checklists, nomination, work-sample tests, free written ratings, behaviorally anchored rating scales, summated rating scales, trait formats, mixed standard scales, and summary scale formats. A few of these show more promise than others and have been evaluated in terms of their ability to reduce rating errors. Some of the more recent formats studied have been the summated rating scale, mixed standard rating scale, behaviorally anchored rating scale and the summary rating scale.

The mixed standard scale was studied by Blanz and Ghiselli (1972). With the mixed standard scale, items relating to each performance dimension are not grouped together but mixed in random order. The purpose is that a rater will not remember how he rated the last item for a particular dimension so it will not affect his rating of the current item. The researchers found that the scale did increase reliability and reduce halo. A later study by Saal and Landy (1977) compared the mixed standard scale with the behaviorally anchored rating

scale and found no advantage for one scale over the other. So its superiority has not been established even though it is logical and appealing.

Another rating scale, the behaviorally anchored rating scale (BARS: Smith and Kendall, 1963) has shown some promise. The BARS format, where each scale point is operationally defined by a statement giving an example of the type of performance behavior, is very appealing to both psychologists and raters. It appears to be a very sound way to develop a rating scale. The only problem is that it has not shown any superiority, psychometrically, over any other rating scale (e.g., Borman, 1979; Dunnette and Borman, 1979; Schwab, Herman, and DeCotiis, 1975).

A third rating scale that has had much support has been the summated rating scale. The summated rating scale uses a set of dimensional items that define different aspects of job performance in terms of behaviors relevant to the job. Each dimensional item is matched with a Likert-type response scale. Most of the research comparing the BARS format with this scale in terms of reduced rater errors (leniency) and reliability favors the summated rating scale (Campbell, Dunnette, Arvey and Hellervik, 1973; Bernardin et al., 1976; Bernardin, 1977). Borman (1979) compared the summated rating scale with BARS format, the summary rating scale, a trait format and number scale and found no format better than

another. From the research it appears there may be limited support for the summated rating scale in reducing rater errors and increasing reliability.

The summary rating scale was defined by Borman (1979) as a modified BARS format where a more general summary anchor was used in place of the highly specific behavioral examples used in the BARS format. The summary rating scale was compared with other rating scales to determine any differences in reliability. The results were that no real difference was found.

From the research on rating scale formats it seems that no one rating has a clear advantage over another in terms of reduced rater errors and increased reliability. This can be complicated further by some research that shows cognitively complex raters may provide better ratings on more complex behavioral rating scales (Schneir, 1977), while less sophisticated raters do better and provide higher quality ratings on simpler rating scales. With all facts considered, the summated rating scale has slightly more support indicating increased reliability and reduced rating errors as compared with other scales.

IV. RATER PERFORMANCE OBSERVATION PROBLEMS

One of the main difficulties for raters trying to rate performance accurately and reliably may be the lack of

opportunity to observe the ratee's performance (Borman, 1979). Thorndike and Hagen found that the accuracy of a rater's rating depended on the opportunity to observe ratee performance. Gunderson and Nelson (1966) studied the intralevel and interlevel ratings of supervisors and peers. The intralevel reliability coefficients were higher than the interlevel reliability coefficients. This supports the contention that more opportunity to observe performance increases reliability because the raters on the same level would have more opportunity to observe each other's performance. It also suggests that rater's rating of performance may interact with a rater's own job and relative level of management. In other research Borman (1974) observed in a field study that interrater reliability increased when scales were used which gave the rater more opportunity to observe performance. Barrett (1966) states that time needed to observe another's performance in order to rate it "depends not only on the complexity of the job, but also on how closely the two work together." Too often a rater is asked to rate a person's performance on a set of rating scales where the rater has only had an opportunity to observe the ratee's performance on a subset of those scales.

V. SUMMARY

Most research on reliability problems have centered on three areas. They are rater errors, rating scale format and

opportunity to observe performance. These reliability problems have been approached in a larger number of studies in an effort to define the problems and offer solutions. The number of studies indicates the importance of seeking high reliability, but warnings of Barrett (1966) and others should be noted. That is that reliability is necessary but not to the exclusion of other measurement requirements such as validity. A reliable but trivial measure is useless.

For the problem of rater errors such as halo, leniency, etc., the solution in most of the research still seems to be training. However, rater errors cannot be separated from rating scale format problems and observation opportunity problems and related solutions. Often other problems such as organizational pressure are considered along with rater errors of leniency. So actually rater errors could be and often are used as the broad heading for all problems of reliability.

No one rating scale format has seemed to emerge as a superior format in terms of reduced rater errors and increased reliability. A great deal of format comparison research ends up showing little difference and often contradictory results. If one format had to be chosen for better reliability the research might support the summated rating scale as a little better. There is some research that indicates reliability of a rating format may be situational (Schneir, 1977).

There seems to be little research on the relationship of opportunity to observe performance and reliability of a performance rating. Most research in this area seems related to the level in organization of rater in relation to the ratee. Then it becomes the problem of separating the effects on reliability of the interaction of rater's relative job level to ratee's with the rating from the interaction of the rater's opportunity to observe performance behavior with the rating. In other words, there is a relationship between relative job level of rater and ratee, and the number and length of opportunities to observe performance. What research that has been done indicates, as one might expect, that reliability increases with the more opportunity there is to observe the performance that is to be evaluated.

CHAPTER III

METHODS AND PROCEDURES

I. SAMPLE

The job titles for which performance ratings were made were District Manager, Relief District Manager, and Utility District Manager as well as some personnel in jobs such as driver, newsstand (rack) service, and single copy personnel for the circulation department. All these jobs had been included in the discussion and development of the questionnaire. The items did not all apply to all jobs but each item applied to some job(s). The primary elements of these jobs included newspaper distribution to local news carriers, maintaining a local carrier force, sales promotion, customer relations and service, record keeping and news rack distribution.

There was one supervisor for each combination of 4 to 10 positions (district managers, utility district managers, relief district managers and single copy personnel) depending on the size and location of the districts. The average size was about eight subordinates to one supervisor. (This supervisor is referred to as the first line supervisor). The first line supervisors were in turn responsible to four department heads and assistants. The number of rating

questionnaires completed by the department heads and their assistants varied widely. One department head completed 62 questionnaires, another completed 32, the third 18 and the fourth completed only 11 questionnaires. In summary there were about 143 persons to be rated for job performance by sixteen first line supervisors and four department heads.

II. INSTRUMENT

The rating instrument was developed for the job performance rating because through use of it a standard set of questions was provided for all ratees by all raters across time. The format for the questionnaire was a behaviorally based summated scale. This format provides a series of statements describing effective job performance behavior and provides a Likert type scale which the rater uses to indicate the degree to which the ratee performs this behavior. The statements were grouped into job dimensions and sub-dimensions. There is some support for this type of rating scale in studies done by Bernardin (1977) and Bernardin et al. (1976). The scale for each statement item had a range of 1 to 7 with 1 being "Superior . . . Performance Superior, usually far exceeds standards and expectations for this district/division" and 7 being "Unacceptable . . . Inferior performance fails to meet standards and expectations for this district/division" (Appendix A).

Another feature of the questionnaire was a confidence rating scale for each statement item. This was a scale with four scale points ranging from "A" to "D." "A" represented a "Very high confidence" with good opportunity to observe and assess performance. "D" represented rare or no opportunity to observe and assess performance. The purpose of this scale was to give the rater a chance to indicate which statements of performance were rated with or without confidence (opportunity to observe). This scale was important for the Department managers who had fewer (or no) opportunities to observe certain behavior than the first line supervisor. It was hypothesized that interrater reliability would increase with higher confidence scores because of less rater error. The items of the questionnaire were grouped into five basic job dimensions: Service, Collections, Sales, Accounting and Record Keeping, and Effort. There are 28 items that relate to service. They are subdivided into groups such as distribution, driving, carrier related service, route organization, etc. The sales dimension is composed of six items grouped into two categories. Four items are included in the collections dimension and six items for accounting and record keeping dimension. The last dimension, effort, has ten items. The total number of items is 54 in this part of the instrument. Included for each of the five dimensions was an overall rating for that dimension.

The last section of the rating questionnaire was a replication of the items in the first four dimensions, excluding the effort dimension. This made a 44 item questionnaire. The only difference was that the seven point scale was used to represent the level of performance predicted for the ratee in future opportunities. Also the confidence scale was dropped because it had little meaning for predicted, unobserved performance.

The seven point prediction scale was given an operable definition so that "1" meant "Superior . . . Learns extremely quickly, expected to far exceed standards and/or expectations promptly" and "7" meant "Inferior . . . Not at all suitable for this work, cannot meet standards or expectations."

III. DATA COLLECTION

The job performance evaluation questionnaires were distributed by the director of personnel to all the first line supervisors and department heads for the distribution department of the newspaper's circulation division. Each was instructed to use the forms as instructed and to fill them out as completely as possible. The department heads and the supervisors were asked to complete a form for each subordinate and to do so independently. It was important to have the raters do independent evaluations in order to make estimates of interrater agreement.

After both the department head and the supervisor had completed their forms independently for each person, the two managers met and completed a third form, discovering and adjusting evaluation differences on the third form only, leaving the two original forms for each employee intact.

The agreed upon job performance evaluation form was transmitted to Circulation Division management for review. After review and approval by top management, the feedback process, not of importance here, began.

There were the usual problems in trying to get two independent job performance evaluations for each ratee in a dynamic, on-going organization. Some men were being hired as others were being terminated. As a result one supervisor would rate an incumbent's performance but the man would be terminated before the other supervisor would make his rating.

The physical distance between some of the districts and main office was also a problem (in some cases more than 80 miles). That reduced the number of ratees on which there were two job performance evaluations made. The only contact with some of job incumbents in more distant locations was through the first line supervisor. Therefore, the department head had no way to observe performance which was necessary in order to complete the rating questionnaire. The combination of these data collection problems reduced the sample size from 143 to 123.

IV. ANALYSIS

Several methods of analysis were used to examine the reliability of the job performance evaluation instrument and process for comparison. To better understand the reliability of the questionnaire as a set of individual items, as subgroups of items, and as a whole, more than one index was used. An overall index may serve for making comparisons with other values reported in the literature, but if just part of the questionnaire were to be used at some future date, reliability of the subdivision scores of the questionnaire would be helpful. Also, it was of interest to note how the reliability indices for the subdivisions of the questionnaire relate to the overall indices.

An item by item correlational analysis was computed using the item scores from the ratings by the first line supervisor and the item scores of the department head for each evaluated employee. The procedure was repeated for the items in the predicted performance section of the questionnaire.

A similar correlational procedure was followed for subgroups of items (total subgroup score) in the questionnaire. The subgroup scores were composed of all items that related to some aspect of the job such as driving, distribution carrier relations, etc. There were ten of these subgroup scores. The items that related to overall performance

on a job dimension were not included in the subgroup scores. An alpha Coefficient (Cronbach, 1972) was computed for each of these ten groups for both the first line supervisor and the department head. This was done so that a measure of internal consistency could be compared with an actual correlation coefficient for the two supervisors' ratings.

This procedure was repeated using nine subgroup scores of predicted performance rating items. The missing tenth group was the "effort" items which were not part of predicted performance section of the questionnaire.

A correlation coefficient was computed for the total score for all items for the first line supervisor's ratings with total score for all items from the department head ratings and for the predicted performance rating section. The alpha coefficient was also computed for both the first line supervisor and the department head for the performance evaluation section and the predicted performance section of the questionnaire taken separately. There was some question as to how much the department head's lack of opportunity to observe job performance behaviors would affect the reliability coefficients in this study. Therefore, the analysis described above was repeated once for only those items on which the department head had 1 or 2 as a confidence rating and then for only those items on which the department head had 3 or 4 as a confidence rating. [That is, the department head

believed that he had at least observed the performance of the questionnaire item with "some frequency" and had a "fair" opportunity to assess performance in the first analysis and he believed that he had observed the performance of the questionnaire item "infrequently" (or never) and had a "poor" (or no) opportunity to assess performance on the second analysis.] Similarly a third analysis was done for only those items on which the department head and first line supervisor had 1 or 2 as a confidence rating for the item. A comparison of these correlations with those correlations made without regard to the rater's confidence in rating performance was made.

There was also a concern that the confidence ratings and the performance ratings would show a relationship. It was suspected that a high rating for one scale could indicate a high rating on the other. A correlation of the two scales was performed for each item, once for department heads, once for first line supervisors and the third time combining the two. Also, overall correlations for the total scores on the questionnaire were performed in the same manner.

CHAPTER IV

RESULTS

The item by item correlations between performance ratings of first line supervisor and department head were moderate. Of the 54 items, 50 had significant correlations ($p = .001$) between .34 and .68. The correlations for the other four items were not significant ($p = .18$) (see Table 1). Three of these items related to automated paper rack sales and the fourth to company vehicle maintenance (for item means and standard deviations, see Table 9, Appendix B).

The item by item correlations for the predicted performance ratings were also moderate. Of the 44 items, 43 of these correlations were significant ($p < .001$) and were between 34 and 73. (See Table 2.) One of the items relating to predicted performance in cleaning mechanical news racks showed no correlation (for item means and standard deviations, see Table 10, Appendix B).

After reviewing the results from the item by item performance evaluation correlations, it was found that a number of items had a very low response rate. It was discovered that this low response rate was due to the fact that the items described a particular job behavior that related to only a few of the ratee's or none of the ratee's. There had been some job reorganization before the questionnaire

TABLE 1
PERFORMANCE ITEM CORRELATIONS OF FIRST LINE
SUPERVISOR RATINGS WITH DEPARTMENT
HEAD RATINGS

Dimension/Subdivision	Item Number	r	N	Level of Significance
<u>Service</u>				
Predistribution	1	.51	122	<.001
	2	.43	117	<.001
	3	.56	123	<.001
Distribution	1	.56	117	<.001
	2	.68	117	<.001
Driving	1	.49	123	<.001
	2	.52	89	<.001
	3	.23	34	<.18*
	4	.34	117	<.001
Carrier related service	1	.53	87	<.001
	2	.63	90	<.001
	3	.66	88	<.001
Personal service on district	1	.52	123	<.001
	2	.52	123	<.001
	3	.54	115	<.001
	4	.63	123	<.001
	5	.59	51	<.001
	6	.44	106	<.001
	7	.64	113	<.001
Route organization	1	.54	76	<.001
Maintains carrier force	1	.59	84	<.001
	2	.55	88	<.001
	3	.53	86	<.001
Maintains racks	1	.12	70	<.34*
	2	.02	70	<.90*
	3	.34	94	<.001
	4	.38	71	<.001
Overall Performance on Service		.37	117	<.001
<u>Sales</u>				
Personal, direct sales	1	.09	90	<.39*
work	2	.58	110	<.001
	3	.45	70	<.001
Carrier-related, indirect	1	.51	83	<.001
sales work	2	.41	87	<.001
Overall Performance on Sales		.54	112	<.001
<u>Collections</u>				
Regular collections	1	.48	85	<.001

TABLE 1 (Cont'd)

Dimension/Subdivision	Item Number	r	N	Level of Significance
Solves collection problems	1	.63	84	<.001
Personal collection work	1	.58	106	<.001
<u>Overall Performance on Collections</u>		.55	.03	<.001
<u>Accounting and Recordkeeping</u>				
	1	.51	122	<.001
	2	.52	100	<.001
	3	.54	106	<.001
	4	.51	95	<.001
	5	.57	82	<.001
<u>Overall Performance on Accounting and Recordkeeping</u>		.49	116	<.001
<u>Effort</u>				
Effort put forth on service	1	.45	90	<.001
	2	.41	121	<.001
Effort put forth on sales related duties	1	.62	117	<.001
Effort put forth on accounting and record keeping duties	1	.61	122	<.001
	2	.53	115	<.001
Effort put forth on collection		.60	111	<.001
Effort put forth on carrier force maintenance		.58	84	<.001
Effort put forth on personal service and route organization		.53	89	<.001
<u>Overall Performance on Effort</u>		.52	120	<.001

*These correlations were not significant but were listed to show the size of the sample.

TABLE 2

PREDICTED PERFORMANCE ITEM CORRELATIONS OF
FIRST LINE SUPERVISOR RATINGS WITH
DEPARTMENT HEAD RATINGS

Dimension/Subdivision	Item Number	r	N
<u>Service</u>			
Predistribution	1	.48*	122
	2	.58*	121
	3	.54*	123
Distribution	1	.60*	123
	2	.63*	122
Driving	1	.43*	123
	2	.49*	102
	3	.45*	55
	4	.42*	116
Carrier related service	1	.72*	98
	2	.64*	98
	3	.61*	98
Personal service on district	1	.61*	123
	2	.62*	123
	3	.61*	121
	4	.60*	123
	5	.62*	65
	6	.52*	117
	7	.55*	120
Route organization		.66*	89
Maintains carrier force	1	.64*	95
	2	.56*	96
	3	.56*	97
Maintain racks	1	.37*	89
	2	--	89
	3	.40*	107
	4	.51*	87
<u>Overall Performance on Service</u>		.36*	120
<u>Sales</u>			
Personal, direct sales work	1	.34*	101
	2	.62*	118
	3	.50*	92
Carrier-related, indirect sales	1	.54*	98
work	2	.56*	99
<u>Overall Performance on Sales</u>		.55*	118
<u>Collections</u>			
Regular collections		.59*	98

TABLE 2 (Cont'd)

Dimension/Subdivision	Item Number	r	N
Solves collection problems		.62*	98
Personal collection work		.56*	111
<u>Overall Performance on Collections</u>		.63*	111
<u>Accounting and Recordkeeping</u>	1	.53*	123
	2	.49*	106
	3	.54*	118
	4	.58*	106
	5	.73*	99
<u>Overall Performance on Accounting and Recordkeeping</u>		.55*	121

*p < .001

was administered. It was for this reason that some items had to be excluded to obtain a sufficient sample size for computation of the overall correlation coefficients for total job performance evaluations and the predicted performance evaluations from the questionnaire. The items that were excluded were: the item that related to company vehicle maintenance (item number 3 under Driving heading in Appendix A), the item that concerned making deliveries of shorts, replacement papers, etc. (item number 5 under Personal service on district heading in Appendix A) and the four items in the section for maintaining racks (items numbered 1-4 under Maintain racks heading in Appendix A.) The resulting overall correlation for the job performance evaluation total scores was .65 with sample size of 59 employees. (Mean = 2.71) The alpha coefficients for performance ratings by the department heads and the first line supervisors were .98 and .97, respectively.

For the predicted performance rating total scores the correlation was .75 with a sample size of 78 employees. The same four items mentioned above were deleted from this analysis also. (Mean = 2.42.) The resulting alpha coefficient for the predicted performance ratings made by department heads was .98 and by the first line supervisors was also .98.

The results of the correlations of job performance evaluation items grouped into ten dimensions shows all

except three were in the .60 to .71 range. (See Table 3.) The items relating to driving had a correlation of .50 with a sample size of 83. This correlation was computed after the item rating company vehicle maintenance was removed to obtain sufficient sample size to perform the correlation (item number 3 under Driving heading in Appendix A). It was found that this item was not applicable to most of the ratees. Another low correlation was obtained for that group of items that related to job performance on service for automated news racks. The correlation was .24 with a sample size of 69 employees. A third low correlation was obtained for those items concerning sales performance. This correlation was .47 with a sample size of 68.

For items of predicted performance grouped in nine job dimension (the effort items were not in the predicted performance part of the questionnaire) the correlations were higher with six dimensions in the range of .66 to .78. (See Table 4.) Again the three dimensions that were lower in the performance evaluation part of the questionnaire were also lower in the predicted performance part of the questionnaire.

The correlation of the overall total score for the first line supervisor rating with the department head rating was .65 with an $N = 97$. The alpha coefficient for the department head was .89 and for the first line supervisor was .89.

TABLE 3
PERFORMANCE ITEM SUBGROUP CORRELATIONS OF FIRST
LINE SUPERVISOR RATINGS WITH DEPARTMENT
HEAD RATINGS

Subdivisions Included	r	N	Alpha Dept. Head	Alpha 1st Line Supr.
Predistribution				
Distribution (Mean=2.17)	.61	111	.91	.89
Driving (Mean=2.35)	.50	83	.71	.68
Carrier related service (Mean=2.85)	.68	86	.91	.90
Personal service on district (Mean=2.41)	.67	105	.82	.81
Route organization				
Maintains carrier force (Mean=2.93)	.65	76	.90	.88
Maintain racks (Mean=2.46)	.24	69	.78	.84
Personal, direct sales				
Carrier-related, indirect sales work (Mean=3.54)	.47	68	.87	.89
Regular collections				
Solves collection problems				
Personal collection work (Mean=2.69)	.63	84	.89	.84
Accounting and record- keeping (Mean=2.22)	.61	81	.90	.90
Effort (Mean=2.69)	.60	83	.92	.91

TABLE 4

PREDICTED PERFORMANCE ITEM SUBGROUP CORRELATIONS
OF FIRST LINE SUPERVISOR RATINGS WITH
DEPARTMENT HEAD RATINGS

Subdivisions Included	r	N	Alpha Dept. Head	Alpha 1st Line Supr.
Predistribution				
Distribution (Mean=1.94)	.68	120	.95	.93
Driving (Mean=2.19)	.73	35	.87	.94
Carrier related service (Mean=2.74)	.72	98	.53	.91
Personal service on district (Mean=2.21)	.78	64	.86	.89
Route organization				
Maintains carrier force (Mean=2.42)	.70	89	.94	.88
Maintain racks (Mean=2.21)	.45	86	.89	.86
Personal, direct sales				
Carrier-related, indirect sales work (Mean=3.19)	.54	89	.95	.88
Regular collections				
Solves collection problems				
Personal collection work (Mean=2.61)	.66	98	.92	.89
Accounting and record- keeping (Mean=2.05)	.67	97	.95	.92

For the predicted performance ratings the correlation of the overall items was .66 with an $N = 108$. The alpha Coefficients were similar to the performance overall item alpha coefficients. For the department head ratings it was .87 and for the first line supervisor it was .87.

When the correlations of items for those ratings where the department head had a high confidence rating were compared with the correlation of items where the confidence rating was unrestricted, the results were not as expected. From the test statistics derived from the comparison of the Z scores for the two sets of correlations, the hypothesis that high department head confidence rating scores indicate an increase in interrater reliability would have to be rejected. It can be seen in Table 5 that none of the interrater correlations where the department had a high confidence rating could be considered higher than the unrestricted correlations under the conditions of a one tailed test at the .05 level. In many cases the correlations for rating scores with high department head confidence rating scores were lower than the correlations of the unrestricted confidence ratings. (Mean confidence rating for department head = 2.47, for supervisor = 1.60.)

The results of the comparison of those item interrater correlations which had a low (3 or 4) department head confidence rating score with the correlations which had

TABLE 5

COMPARISON OF ITEM CORRELATIONS HAVING HIGH
DEPARTMENT HEAD CONFIDENCE RATING WITH
ITEM CORRELATIONS HAVING UNRESTRICTED
CONFIDENCE RATING

Dimension/Subdivision	Item Number	r Unrestr. Conf.	N	r High Conf.	N	Standard Score Computed from Z
<u>Service</u>						
Predistribution	1	.51**	122	.46**	82	.47
	2	.43**	117	.30**	70	1.00
	3	.56**	123	.63**	61	-.62
Distribution	1	.56**	117	.46**	48	.75
	2	.68**	117	.56**	46	1.13
Driving	1	.49**	123	.54**	55	-.43
	2	.52**	89	.61**	50	-.72
	3	.23	34	.79	3	-.81
	4	.34**	117	.00	29	1.61
Carrier related service	1	.53**	87	.53**	41	.04
	2	.63**	90	.67**	42	-.38
	3	.66**	88	.68**	40	-.15
Personal service on district	1	.52**	123	.53**	99	-.10
	2	.52**	123	.50**	84	.17
	3	.54**	115	.33	29	1.23
	4	.63**	123	.63**	102	.08
	5	.59**	51	.54**	47	.38
	6	.44**	106	.48**	49	-.22
	7	.64**	113	.66**	59	-.20
Route organization	1	.54**	76	.51**	38	.17
Maintains carrier force	1	.59**	84	.61**	47	-.13
	2	.55**	88	.58**	33	-.24
	3	.53**	86	.57**	51	-.29
Maintain racks	1	.12	70	-.08	27	.81
	2	.02	70	.01	25	.02
	3	.34**	94	.15	32	.97
	4	.38**	71	.07	33	1.52
<u>Overall Performance on Service</u>		.37**	117	.31*	44	.33
<u>Sales</u>						
Personal, direct sales work	1	.09	90	-.02	39	.58
	2	.58**	110	.58**	29	-.05
	3	.45**	70	.57*	25	-.65
Carrier-related, indirect sales work	1	.51**	83	.62**	36	-.83
	2	.41**	87	.29	31	.63

TABLE 5 (Cont'd)

Dimension/Subdivision	Item Number	r Unrestr. Conf.	N	r High Conf.	N	Standard Score Computed from Z
<u>Overall Performance</u>						
<u>on Sales</u>		.54**	112	.51*	33	.24
<u>Collections</u>						
Regular collections	1	.48**	85	.53**	34	-.35
Solves collection problems	1	.63**	84	.53*	30	.67
Personal collection work	1	.58**	106	.52**	45	.41
<u>Overall Performance</u>						
<u>on Collections</u>		.55**	103	.59**	41	-.27
<u>Accounting and</u>						
<u>Recordkeeping</u>						
	1	.51**	122	.48**	88	.32
	2	.52**	100	.57**	62	-.41
	3	.54**	106	.52**	68	.15
	4	.51**	95	.50**	76	.02
	5	.57**	82	.61**	46	-.36
<u>Overall Performance</u>						
<u>on Recordkeeping</u>		.49**	116	.54**	55	-.42
<u>Effort</u>						
Effort on service	1	.45**	90	.37**	38	.48
	2	.41**	121	.43**	53	-.15
Effort on sales	1	.62**	117	.58**	58	.40
	2	.44**	88	.41**	35	.18
Effort on record-	1	.61**	122	.59**	77	.31
keeping duties	2	.53**	115	.50**	37	.19
Effort on collection	1	.60**	111	.55**	62	.42
Effort on carrier	1	.58**	84	.55**	48	.20
force maintenance						
Effort on personal	1	.53**	89	.59**	49	-.49
<u>Overall Performance</u>						
<u>on Effort</u>		.52**	120	.50**	51	.23

*p < .05

**p < .001

unrestricted confidence rating scores showed again no difference. From Table 6 it can be seen that the hypothesis that correlations with a lower department head confidence rating scores would be lower than those correlations with an unrestricted confidence rating score was rejected. And in many cases the expected lower correlations turned out to be higher.

In Table 7 the comparison of item interrater correlations which had high department head confidence rating scores with item interrater correlations which had low department head confidence rating scores had similar results as the two previous analyses. And from Table 8 where a comparison was made between item interrater correlations which had high confidence rating scores from both raters and item interrater correlations which had unrestricted confidence rating scores, the results were comparable. There seemed to be no support for hypothesis that a higher or lower confidence rating score meant a higher or lower interrater reliability respectively.

In examining the results of the correlational analyses between confidence rating scores and performance rating scores within and across raters it was found that there were few significant correlations. The few correlations that did reach a level of significance could still be explained by chance because of the number of correlations calculated.

TABLE 6

COMPARISON OF ITEM CORRELATIONS HAVING LOW
DEPARTMENT HEAD CONFIDENCE RATING WITH
ITEM CORRELATIONS HAVING UNRESTRICTED
CONFIDENCE RATING

Dimension/Subdivision	Item Number	r Unrestr. Conf.	N	r Low Conf.	N	Standard Score Computed from Z
<u>Service</u>						
Predistribution	1	.51**	122	.59**	31	-.59
	2	.43**	117	.54**	41	-.72
	3	.56**	123	.37*	52	1.40
Distribution	1	.56**	117	.55**	69	.13
	2	.68**	117	.70**	71	-.22
Driving	1	.49**	123	.36*	61	1.02
	2	.52**	89	.34*	37	1.12
	3	.23	34	.42	26	-.78
	4	.34**	117	.48**	82	-1.15
Carrier related service	1	.53**	87	.58**	46	-.35
	2	.63**	90	.61**	48	.16
	3	.66**	88	.67**	48	-.10
Personal service on district	1	.52**	123	.41	18	.51
	2	.52**	123	.60**	33	-.60
	3	.54**	115	.56**	86	-.18
	4	.63**	123	-.1*	15	.14
	5	.59**	51	.90	4	-.80
	6	.44**	106	.41*	56	.28
	7	.64**	113	.57**	54	.65
Route organization	1	.54**	76	.56**	38	-.17
Maintains carrier force	1	.59**	84	.58**	37	.08
	2	.55**	88	.55**	54	.04
	3	.53**	86	.48*	34	.33
Maintain racks	1	.12	70	.38*	34	-1.30
	2	.02	70	.20	36	-.89
	3	.34**	94	.31*	54	.16
	4	.38**	71	.64**	30	-1.57
<u>Overall Performance on Service</u>		.37**	117	.45**	63	-.60
<u>Sales</u>						
Personal, direct sales work	1	.09	90	.21	45	-.62
	2	.58**	110	.57**	79	.08
	3	.45**	70	.36*	45	.53

TABLE 6 (Cont'd)

Dimension/Subdivision	Item Number	r Unrestr. Conf.	N	r Low Conf.	N	Standard Score Computed from Z
Carrier-related, indirect sales work	1	.51**	83	.38*	47	.81
	2	.41**	87	.56**	56	-1.16
<u>Overall Performance on Sales</u>		.54**	112	.55**	68	.30
<u>Collections</u>						
Regular collections	1	.48**	85	.42*	51	.39
Solves collection problems	1	.63**	84	.67**	54	-.37
Personal collection work	1	.58**	106	.64**	54	-.59
<u>Overall Performance on Collections</u>		.55**	103	.55**	56	.04
<u>Accounting and Recordkeeping</u>	1	.51**	122	.61**	28	-.61
	2	.52**	100	.43*	33	.56
	3	.54**	106	.56**	35	-.14
	4	.51**	95	.47	13	.13
	5	.57**	82	.41*	32	.96
<u>Overall Performance on Recordkeeping</u>		.49**	116	.48**	51	.12
<u>Effort</u>						
Effort on service	1	.45**	90	.53**	51	-.64
	2	.41**	121	.36*	61	.32
Effort on sales	1	.62**	117	.64**	51	-.14
	2	.44**	88	.49**	51	-.38
Effort on record-keeping duties	1	.61**	122	.73**	38	-1.07
	2	.53**	115	.58**	71	-.50
Effort on collection	1	.60**	111	.69**	42	-.84
Effort on carrier force maintenance	1	.58**	84	.58**	35	-.05
Effort on personal	1	.53**	89	.39*	38	.85
<u>Overall Performance on Effort</u>		.52**	120	.56**	62	-.32

*p < .05

**p < .001

TABLE 7

COMPARISON OF ITEM CORRELATIONS HAVING HIGH
DEPARTMENT HEAD CONFIDENCE RATING WITH
ITEM CORRELATIONS HAVING LOW DEPART-
MENT HEAD CONFIDENCE RATING

Dimension/Subdivision	Item Number	r Low Conf.	N	r High Conf.	N	Standard Score Computed from Z
<u>Service</u>						
Predistribution	1	.59**	31	.46**	82	.87
	2	.54**	41	.30**	70	1.42
	3	.37*	52	.63**	61	-1.81***
Distribution	1	.55**	69	.46**	48	.58
	2	.70**	71	.56**	46	1.21
Driving	1	.36*	61	.54**	55	-1.23
	2	.34*	37	.61**	50	-1.59
	3	.42*	26	.79	3	1.48
	4	.48**	82	.00	29	2.29
Carrier related service	1	.58**	46	.53**	41	.33
	2	.61**	48	.67**	42	-.46
	3	.67**	48	.68**	40	-.05
Personal service on district	1	.41	18	.53**	99	-.55
	2	.60**	33	.50**	84	.69
	3	.56**	86	.33	29	1.31
	4	.61*	15	.63**	102	-.10
	5	.90	4	.54**	47	.88
	6	.41*	56	.48**	49	-.43
Route organization	1	.57**	54	.66**	59	-.73
	2	.56**	38	.51**	38	.29
Maintains carrier force	1	.58**	37	.61**	47	-.17
	2	.55**	54	.58**	33	-.25
	3	.48*	34	.57**	51	-.53
Maintain racks	1	.38*	34	-.08	27	1.74
	2	.20	36	.01	25	.71
	3	.31*	54	.15	32	.78
	4	.64**	30	.07	33	2.60
<u>Overall Performance on Service</u>		.45**	63	.31*	44	.77
<u>Sales</u>						
Personal, direct sales work	1	.21	45	-.02	39	1.02
	2	.57**	79	.58**	29	-.10
	3	.36*	45	.57*	25	-1.01

TABLE 7 (Cont'd)

Dimension/Subdivision	Item Number	r Low Conf.	N	r High Conf.	N	Standard Score Computed from Z
Carrier-related, indirect sales work	1	.38*	47	.62**	36	-1.41
	2	.56**	56	.29	31	1.46
<u>Overall Performance on Sales</u>		.55**	68	.51*	33	.31
<u>Collections</u>						
Regular collections	1	.42*	51	.53**	34	-.63
Solves collection problems	1	.67**	54	.53*	30	.90
Personal collection work	1	.64**	54	.52**	45	.86
<u>Overall Performance on Collections</u>		.55**	56	.59**	41	-.28
Accounting and Recordkeeping	1	.61**	28	.48**	88	.79
	2	.43*	33	.57**	62	-.82
	3	.56**	35	.52**	68	.24
	4	.47	13	.50**	76	-.12
	5	.41*	32	.61**	46	-1.15
<u>Overall Performance on Recordkeeping</u>		.48**	51	.54**	55	-.45
<u>Effort</u>						
Effort on service	1	.53**	51	.37**	38	.95
	2	.36*	61	.43**	53	-.40
Effort on sales	1	.64**	51	.58**	58	.45
	2	.49**	51	.41**	35	.46
Effort on record-keeping duties	1	.73**	38	.59**	77	1.22
	2	.58**	71	.50**	37	.54
Effort on collection	1	.69**	42	.55**	62	1.09
Effort on carrier force maintenance	1	.58**	35	.55**	48	.21
Effort on personal	1	.39*	38	.59**	49	-1.16
<u>Overall Performance on Effort</u>		.56**	62	.50**	51	.46

*p < .05

**p < .001

***p < .05 for a one tailed test

TABLE 8

COMPARISON OF ITEM CORRELATIONS HAVING HIGH
DEPARTMENT HEAD AND SUPERVISOR CONFIDENCE
RATINGS WITH ITEM CORRELATIONS HAVING
UNRESTRICTED CONFIDENCE RATINGS

Dimension/Subdivision	Item Number	r Unrestr. Conf.	N	r High Conf.	N	Standard Score Computed from Z
<u>Service</u>						
Predistribution	1	.51**	122	.43**	78	.69
	2	.43**	117	.27*	67	1.22
	3	.56**	123	.57**	56	-.09
Distribution	1	.56**	117	.49**	45	.55
	2	.68**	117	.53**	43	1.26
Driving	1	.49**	123	.53**	49	-.36
	2	.52**	89	.58**	47	-.42
	3	.23	34	--	--	--
	4	.34**	117	-.01	26	1.57
Carrier related	1	.53**	87	.52**	38	.09
	2	.63**	90	.66**	38	-.24
	3	.66**	88	.67**	38	-.10
Personal service on district	1	.52**	123	.53**	95	-.09
	2	.52**	123	.53**	81	-.06
	3	.54**	115	.23	25	1.60
	4	.63**	123	.66**	96	-.38
	5	.59**	51	.51**	44	.58
	6	.44**	106	.52**	48	-.54
	7	.64**	113	.66**	56	-.20
Route organization	1	.54**	76	.53*	34	.09
Maintains carrier	1	.59**	84	.61**	43	-.10
	2	.55**	88	.59**	30	-.23
	3	.53**	86	.56**	49	-.20
Maintain racks	1	.12	70	-.05	26	.71
	2	.02	70	.02	24	.00
	4	.38**	71	.05	32	1.58
<u>Overall Performance on Service</u>		.37**	117	.30	40	.21
<u>Sales</u>						
Personal, direct sales work	1	.09	90	-.03	34	.60
	2	.58**	110	.59*	23	-.07
	3	.45**	70	.53*	20	-.40
Carrier-related, indirect sales work	1	.51**	83	.68**	27	-1.13
	2	.41**	85	.27	26	.64

TABLE 8 (Cont'd)

Dimension/Subdivision	Item Number	r Unrestr. Conf.	N	r High Conf.	N	Standard Score Computed from Z
<u>Overall Performance on Sales</u>		.54**	112	.52*	31	.11
<u>Collections</u>						
Regular collections	1	.48**	85	.52*	32	-.27
Solves collection problems	1	.63**	84	.52*	28	.74
Personal collection work	1	.58**	106	.55**	44	.22
<u>Overall Performance on Collections</u>		.55**	103	.64**	30	-.64
<u>Accounting and Recordkeeping</u>						
	1	.51**	122	.46**	84	.51
	2	.52**	100	.56**	57	-.34
	3	.54**	106	.49**	63	.40
	4	.51**	95	.50**	73	
	5	.57**	82	.57**	42	-.06
<u>Overall Performance on Recordkeeping</u>		.49**	116	.51**	44	-.12
<u>Effort</u>						
Effort on service	1	.45**	90	.35*	35	.57
	2	.41**	121	.44*	50	-.26
Effort on sales	1	.62**	117	.59**	51	.26
	2	.44**	88	.42*	32	.12
Effort on record-keeping duties	1	.61**	122	.59**	72	.30
	2	.53**	115	.46*	33	.46
Effort on collection	1	.60**	111	.54**	56	.47
Effort on carrier force maintenance	1	.58**	84	.53**	45	.34
<u>Overall Performance on Effort</u>		.52**	120	.48*	42	.34

*p < .05

**p < .001

The results can be reviewed in Tables 13, 14, and 15 in Appendix C. The correlations between confidence rating scores and the performance rating scores overall could not be computed because the sample size was too small. There appears to be no correlation between confidence rating scores and performance rating scores in general.

CHAPTER V

DISCUSSION OF RESULTS

As would be expected, single item reliability coefficients are generally lower than reliability coefficients for the subgroup scores and total scores. The item inter-rater reliability coefficients did not vary greatly except for job performance items involving newsrack maintenance and sales, and driving. Most of the item reliability coefficients values center around .54. The item reliability coefficients for the scores of the overall items for each subgroup were in the range of .49 to .54 except for overall performance on service which was .37. That lower value (.37) could have resulted because of the larger number of heterogeneous job performance behaviors it included as compared with the other overall job performance items. There were 27 job performance behavior items to be considered in the service overall rating as opposed to 9 items in the next largest job dimension, effort. It would be more likely with the larger number job performance behavioral items to consider and with differential opportunities to observe different categories of behavior on the job that rater errors caused by failure to observe performance behavior would increase, thus increasing the rater disagreement for the subgroup score.

The predicted job performance item interrater reliability coefficients were slightly larger than the job performance item reliability coefficients. Also the sample size of the number of employees rated by both the department head and the first line supervisor was larger for items of predicted job performance than job performance except for one item. There is no clear indication of cause for this slight increase in reliability coefficients. It could be that it is an artifact of "halo" of some global perception of the employee shared by the department head and the first line supervisor.

The interrater reliability coefficient for each subgroup of items of job performance rating scores was in the range of .47 to .71 except for the subgroup of items for rack maintenance and sales. Excluding the exceptions, the interrater reliability coefficients are approximately the same magnitude as those from several other studies using two different levels of management as raters (Lawler, 1967; Zedeck and Baker, 1972; Campbell et al., 1973). The interrater reliability coefficient that was quite low (.24) was probably caused by lack of opportunity to observe job performance behavior by the department heads. It was reported that department heads rarely, if ever, inspected the racks. Again, the predicted interrater reliability coefficients were slightly higher for the subgroup scores.

The alpha coefficients for the subgroups for each rater were high, in the .80 to .90 range, indicating good

internal consistency within the subgroups. The exceptions were the subgroups relating the job performance on driving and the department heads' performance rating scores on items for maintaining racks. The lower alpha coefficient (.78) for department heads rating scores on items relating to maintaining racks would support the report that department heads rarely observed job performance in this area. The alpha coefficients for both the department heads and first line supervisors were lower, .71 and .68 respectively, for the groups of items that measured job performance on driving. The lower internal consistency of this group for both supervisors would indicate this may be more than just lack of opportunity to observe performance. It could mean some of the items in the subgroups did not measure job performance on driving for the jobs measured. This concern is supported by the fact that one of the items for driving job performance had to be dropped from the analysis earlier because the interrater reliability coefficient for that item was not significant and the number of raters that rated that item was small. The reason for all these inconsistencies was that the company was in the process of separating driving functions of the jobs measured and creating a new job description as driver.

The alpha coefficients for predicted performance rating scores were higher, .85 to .95, than the performance rating

scores except for one. The exception was the subgroup of items that measured predicted performance on carrier related service as rated by the department heads (see Table 4, page 41). This lower alpha coefficient of .53 would indicate that the department heads when predicting performance on carrier related service saw the three items as independent of each other in the general dimension of carrier related service. But when the alpha coefficient of the subgroup of items for carrier related service for the supervisors was examined the high score of .91 indicated actual observation of the job performance tended to show the three items did have some common factor or relation. Thus the internal consistency for a subgroup of items may be perceived differently by different raters. Another explanation might be halo error in the first line supervisors' performance rating.

The overall interrater reliability coefficient for the performance rating part of the questionnaire taken as a whole was reasonably high when compared with other performance evaluation studies in which the interrater reliabilities ranged from .07 to .61 (Zedek and Baker, 1972; Campbell et al., 1973; Dickinson and Tice, 1973; William and Seiler, 1973; Borman, 1974; Borman and Vallon, 1974; Arvey and Hoyle, 1974). The alpha coefficients for department head and first line supervisor performance ratings taken as a whole

were about 30 points higher than the interrater reliability coefficient for the performance ratings overall. This difference, as noted by Nunnally (1967), would indicate some source of error other than the error in "sampling of content." This difference could be attributed to the difference in organizational levels of the raters and differences in the opportunity to observe job performance behavior. The very high alpha coefficients show that content sampling error was very low when usually, according to Nunnally (1967), content sampling error is a major source of measurement error. The fact that alpha coefficients for both raters were equally high would reinforce the idea that the performance rating questionnaire has strong internal consistency.

Of course these high alpha coefficients could indicate a large "halo" rater error effect in the performance ratings. This would mean that the raters had some overall perception of the ratees' performance and allowed this perception to influence their performance ratings on individual items. This would be an undesirable situation for a performance rating. But there are other possible causes for this high internal consistency. One possible alternative reason is that there is a time constraint of job performance so it is impossible for a ratee to have a superior performance on every job performance rating item and still perform the

complete job in an outstanding manner. In this case it would seem unfair for the rater to rate a job incumbent low on some items which were performed in an adequate manner under the time constraints of the job if the ratee performed the overall job in a superior manner. One possible relief from these sources of pressure for uniform ratings would be the inclusion of an overall job performance rating item in the rating format to let the rater "get it out of their system" as suggested in a paper by Whitlock as reported by Lawler (1967).

Another explanation for the high alpha coefficients for the whole rating questionnaire is that there are relationships between how well a person performs on job behavior and how well he performs another. This was one of the possible reasons for halo in the discussion by Barrett (1967).

The interrater reliability coefficient for the overall items taken as a group was equal to the overall interrater reliability coefficient which supports the idea that the performance rating questionnaire can be used with an adequate degree of accuracy. The alpha coefficients for the combined overall items for each group of raters were high but not quite as high as the overall alpha coefficients. This small difference was probably the result of the difference in the number of items in the analysis.

The comparison of the interrater reliability coefficients for those items with a high confidence rating by the

department head with the interrater reliability coefficients for the items with no restriction on the confidence rating showed no improvement in reliability with high department head confidence ratings. Also when the interrater reliability coefficients for those items with low confidence ratings by the department head were compared with the interrater reliability coefficients for items with unrestricted confidence ratings no decrease in reliability was indicated with lower department head confidence rating. In the same manner when the interrater reliability coefficients for items with low department head confidence ratings were compared with reliability coefficients for items with high department head confidence ratings no improvement was shown in reliability for high department head confidence ratings over low department head confidence ratings. In the fourth analysis of a similar nature interrater reliability coefficients for items with high confidence ratings by both department head and first line supervisor were compared with interrater reliability coefficients with unrestricted confidence ratings showed no improvement in reliability for items with high confidence ratings. From these results it can only be concluded that in using this format the raters were not able to report their opportunity to observe job performance behavior in a meaningful way that was reflected in the interrater reliability coefficients for those items.

There may be some restrictions on the results using the ratios computed from the Z scores of correlations because of the smaller sample sizes. Also because the same subjects were included in the unrestricted confidence ratings for three of the four analyses the independence of the two samples was jeopardized which actually would increase the probability of the hypothesis of no difference, a very conservative stance. The failure of the comparison of interrater reliability coefficients for items with high department head confidence ratings vs. low department head confidence ratings to show any difference, would indicate that the lack of independence of the two samples had no meaningful effect. In this case confidence rating contributed nothing to the process while adding quite some time and effort to the rating task.

From the results of the correlation analysis of performance ratings with confidence ratings for each item it appears that most of the items do not show a correlation that is significant. While it was expected that there would be a definite correlation between performance ratings and confidence ratings similar to other studies wherein has been found a correlation between performance rating and item importance (Larsen, 1963), no definite correlation was found here. This would indicate that the raters were rating something with the confidence scale other than performance,

.

but from the results the confidence scale does not seem to make any contribution to reliability.

There were a few items on the job performance rating questionnaire that had correlations with the confidence ratings that reached a level of significance. Although these correlations could be explained as chance occurrences because of the number of items in the analysis, there are some patterns that lead to other explanations. One such pattern is that of the first five items of the rating form which deal with distribution and predistribution. These items of job performance were items that were easily observed and therefore had higher overall confidence ratings. Since these items of job behavior were easily observed by the "boss" it was easy to see that the ratee would have made an extra effort to do these jobs well, therefore there would be a correlation between the two. There was another group of items for which similar reasoning applies; they relate to job performance on personal service to the district. Here again the job performance was easily observed. One of these items showed a correlation between job performance and confidence rating for only the first line supervisor. There was an obvious reason for this since the job performance item dealt with checking in with the supervisor. Another group of items relating to job performance with high visibility for the supervisor was composed of record keeping

items. The items correlated with confidence ratings for only the supervisor since the supervisor is the one inspecting the job incumbents records on a regular basis. One item of job performance had a negative correlation with confidence rating for both supervisor and department head. When it came to keeping racks clean it seem the only time the department head knew about it was when there was a negative report. This is an indication of lack of opportunity to observe performance on rack maintenance by the department head.

In summarizing the discussion on confidence rating results, it seems that the confidence ratings had little relation to reliability and only about a third of the item correlations between confidence rating and performance ratings showed significant correlations.

CHAPTER VI

SUMMARY, CONCLUSION, AND IMPLICATIONS

I. SUMMARY

This study was used to examine the reliability of a performance rating questionnaire in a typical on-the-job environment. The supervisors and department heads of the circulation division of large newspaper joined with a consultant to create a job performance rating form for their division. After several meetings to refine and rewrite the form, a final version acceptable to all, was completed. The rating form used a summated rating format with a seven point Likert type scale and was divided in several groups and subgroups of items. Also part of each item was a four point Likert type scale for rating the confidence of the rater in observing the job performance behavior for that item. There was a second section of the rating form similar to the first for rating predicted performance.

The raters, already familiar with the form since they helped to write it, were trained in the use of the form and some of the problems associated with the rating of job performance. The job incumbents were then rated by their supervisor and department head using the job performance rating form. Some of the problems recognized as detriments

to good reliability were time differentials between the two ratings of an individual's performance, an uncertain time span for observing ratee performance on which the ratings were based, varying numbers and lengths of opportunities to observe performance, the use of raters from two different levels of management for each performance rating and the use of one inclusive form to rate jobs from more than just the three major job titles.

A part of this study was the effort to examine the use of a second scale, one measuring a rater's confidence in his rating based on the rater's opportunity to observe job performance behavior. It was hypothesized that a higher confidence rating score for the department head would be reflected in higher reliability coefficients for those items because the rater error associated with the lack of opportunity to observe would be less.

The reliability of the performance ratings made using the rating questionnaire were analyzed in several ways. The interrater reliability coefficients between the department head and first line supervisor for each item of the job performance rating and the predicted performance rating were computed. The job performance rating items and predicted job performance items were grouped into subgroups and interrater reliability coefficients for these subgroups of items were computed as well as the alpha coefficients for each group of

items for each rater. An overall interrater reliability coefficient for all the job performance rating items taken together and all the predicted job performance rating items taken together (excluding overall items) was computed. The alpha coefficients for each rater level were also determined for the overall groups of job performance ratings and predicted performance ratings. The overall items for job performance ratings and predicted performance were grouped together to be analyzed in the same manner as the previous analyses.

To analyze the relationship of confidence rating scores to the interrater reliability coefficients four analyses were performed. The interrater reliability coefficients for those items with high department head confidence ratings were compared to the interrater reliability coefficients for some items with no restriction on the confidence ratings. This comparison was repeated except the interrater reliabilities for items with low confidence ratings by the department head were used in place of the high confidence ratings. Another comparison was made similar to the first except the interrater reliability coefficients with high department head confidence ratings also were required to have high first line supervisor confidence ratings. In the fourth comparison the interrater reliability coefficients for those items with high department head confidence ratings were compared with

the interrater reliability coefficients for the same items with low department head confidence ratings.

The last part of the analysis consisted of a correlational analysis of job performance rating items and corresponding confidence ratings to determine the relationship between confidence ratings and job performance ratings. This analysis was repeated three times, once for performance ratings of the department head, once for the ratings of the first line supervisor, and a third time for both groups combined.

The interrater reliability coefficients from the first part of the analysis were generally moderate for items as well as for overall and for subgroups. There were a few low interrater reliability indices for subgroups and individual items that were anticipated because of relatively few jobs to which these items applied. The alpha coefficients were generally very high, ranging from .85 to .98. The results from the second part of the analysis clearly indicated no difference in interrater reliability coefficients for items with high or low department head confidence ratings and the interrater reliability coefficients with unrestricted confidence ratings. Also the reliability coefficients seemed no different for job performance ratings with high department head confidence ratings. The correlation analysis for job performance ratings with the confidence ratings reveal no

general pattern of correlations. There were a few correlations between the job performance ratings and confidence ratings that were of interest.

II. CONCLUSIONS

The reliability of this job performance evaluation method was high in comparison with those found in other field studies (on the basis of the interrater reliability coefficients). Both overall and dimension or subgroups of items had interrater reliability coefficients that were as good or better than other field studies using similar methods. This shows that with proper attention to the development of a rating form by active involvement of the users and reasonable training for the raters, a job performance rating questionnaire using summated format can be developed that will have adequate reliability.

The high reliability of the performance rating form was not only indicated by the overall total interrater reliability coefficients, but also in the interrater reliability coefficients for the subgroups of items and the individual items, in most cases. It was not expected to find item reliability coefficients this high in a field setting where raters were from two different levels of management. These reliability coefficients compare favorably with some controlled laboratory studies (Borman, 1978).

These reasonably high interrater reliability coefficients suggest that adequate levels reliability can be obtained for a performance rating form in a field setting where the raters are involved in the creation of the form, under careful guidance. By involving the raters in the creation of the performance evaluation form it gives the form meaning in terms the rater understands which in turn should increase reliability, as in this case. Another conclusion is that rater involvement in the development of the rating form increased rater identity with the form and thus a desire to see it work well. As Taft (1955) said "motivation to make good ratings may be the single most important factor in increasing reliability." Again it must be emphasized that from the interrater reliability coefficients in this study it was shown that reasonably high reliability can be obtained for a performance evaluation form in a field setting.

It cannot be concluded that this job performance evaluation was free from all rater errors. The very high alpha coefficients for job performance ratings for both department heads and first line supervisors overall indicated a high internal consistency. If this high internal consistency is seen negatively as "halo" it could mean that the raters had a global rating for each ratee and forced their rating on each item of job performance to match that global rating. Or it may be that most of the items are

important and form one large dimension of performance instead of several smaller independent dimensions. Another possibility would follow what Barrett (1966) suggested that people are consistent in their behavior across several dimensions. It is therefore not clear what type of connotation should be placed on this indication of high internal consistency.

In examining the results of the comparisons of interrater reliability coefficients for high and low confidence ratings it cannot be concluded that there is any relationship. The hypothesis that any rater error caused by the department head's lack of opportunity to observe job performance behavior would be measured by the interrater reliability coefficient and the level indicated by the confidence rating was rejected. It appears that the confidence rating as it was used here has no use as a measure of the error in performance rating caused by the lack of opportunity to observe job performance behavior. It can also be concluded that there was no correlation between the job performance ratings and the confidence ratings. That means the raters were able to discern that two different scales for performance ratings and confidence ratings measured different things but it is not clear what the raters were measuring with the confidence scales.

III. IMPLICATIONS

From the interrater reliability in this study and others it seems that this may be the best that can be expected in an on-the-job situation with the methods available at the present. It appears that in most organizations, where it is not likely to find two raters on the same supervisory level and with the same opportunities to observe performance (Gunderson and Nelson, 1966), the interrater reliability will not improve.

In this study it was also noted that there is no clear understanding of the presence of high internal consistency in a performance measure. It would be of interest to replicate this study (in process) and include an item that could be used to measure total, overall performance. It could then be seen if this relieved the pressure on the rater to express this general impression of the ratee as suggested by Whitlock as reported by Lawler (1967) and change the internal consistency of the performance ratings. Also the question of how time constraints of a job affect the internal consistency of a performance ratings needs to be investigated. That is, if a ratee performs certain elements or items of job performance at less than superior performance because of time limitations caused by performing other parts of job, will change the way the rater rates those elements or items? These are a few of the questions that need to be

asked in order to add to the understanding of the complex cognitive process used by raters.

Another implication from this study would be the contraindicated use of the confidence scale as an indicator of the rater error caused by the lack of opportunity to observe job performance. It should also be noted that there was no correlation of performance ratings with confidence ratings which raises the question of what the confidence rating is measuring. It is possible that the raters saw the confidence rating, in some cases, as a measure of their own performances as supervisor or department head although no evidence to support that suspicion is reported here. If the raters equated low confidence with not supervising properly then the confidence ratings would be a self report of performances by the raters.

From this research and other studies it seems there is a need for more research in understanding the performance rating process. It seems that the best method of performance evaluation depends on the situation in which it is used. In this case it was demonstrated that careful planning, participation of raters in developing this form, rater training and, at least, quasi independent measure by two levels of supervision resulted in a usefully reliable and usable job performance evaluation form.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Anastasi, A. Differential Psychology. 3rd ed. Macmillan, 1958.
- Anderson, H. E., S. L. Roush, and J. E. McClary. Relationships among ratings, production efficiency, and the general aptitude test battery scales in an industrial setting. Journal of Applied Psychology, 1973, Vol. 58, 77-82.
- Arvey, R. D. and J. C. Hoyle. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 1974, Vol. 59, 61-68.
- Asch, S. E. Forming impressions of personality. Journal of Abnormal and Social Psychology, 1946, Vol. 41, 258-290.
- Barrett, Richard S. Performance Rating. Chicago: Science Research Associates, Inc., 1966.
- Barrett, R. S., E. K. Taylor, J. W. Parker, and L. Martens. Rating scale content I: Scale information and supervisory ratings. Personnel Psychology, 1958, Vol. 11, 333-346.
- Bass, B. M. Reducing leniency in merit ratings. Personnel Psychology, 1956, Vol. 9, 359-369.
- Bernardin, H. J. Behavioral expectation scales vs. summated scales: A fairer comparison. Journal of Applied Psychology, 1977, Vol. 62, 422-428.
- Bernardin, H. J. The effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 1978, Vol. 63, 301-308.
- Bernardin, H. J., K. M. Alvares, and C. J. Cranny. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 1976, Vol. 61, 564-570.
- Bernardin, H. J., and C. S. Walter. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, Vol. 62, 64-69.

- Bingham, W. V. Halo invalid and valid. Journal of Applied Psychology, 1939, Vol. 23, 221-228.
- Bittner, R. Developing an industrial merit rating procedure. Personnel Psychology, 1948, Vol. 1, 403-432.
- Blanz, F., and E. E. Ghiselli. The mixed standard scale: a new rating system. Personnel Psychology, 1972, Vol. 25, 185-199.
- Borman, W. C. The ratings of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 1974, Vol. 12, 105-124.
- Borman, W. C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology. 1975, Vol. 60, 556-560.
- Borman, W. C. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, Vol. 63, 135-144.
- Borman, Walter C. Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 1979, Vol. 64, 410-421.
- Borman, W. C. and W. R. Vallon. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology, 1974, Vol. 59, 197-201.
- Brown, E. M. Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 1968, Vol. 52, 195-199.
- Burnaska, R. F. and T. D. Hollman. An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 1974, Vol. 59, 307-312.
- Campbell, J. P., M. D. Dunnett, E. E. Lawler, and K. E. Weick. Managerial Behavior, Performance, and Effectiveness. New York, N. Y.: McGraw-Hill, 1970.
- Campbell, John P., Marvin D. Dunnette, Richard D. Arvey, and Lowell V. Hellervik. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, Vol. 57, 15-22.

- Cronbach, Lee J. and Goldine C. Gleser. Psychological Tests and Personnel Decisions. Urbana, Ill.: University of Illinois Press, 1965.
- Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. The Dependability of Behavioral Measurements. New York, N. Y.: John Wiley and Sons, Inc., 1972.
- Crow, W. J. The effect of training upon accuracy and variability in interpersonal perception. Journal of Abnormal and Social Psychology, 1957, Vol. 55, 355-359.
- Cummings, L. L. and Donald P. Schwab. Performance in Organizations. Glenview, Ill.: Scott, Foresman and Company, 1973.
- Dickinson, T. L. and T. E. Tice. A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 1973, Vol. 9, 421-438.
- Dunnette, M. D. and W. C. Borman. Personnel selection and classification systems. Annual Review of Psychology, 1979, Vol. 30, 477-525.
- Evans, J. W. Emotional bias in merit rating. Personnel Journal, 1950, Vol. 28, 290-291.
- Fisher, R. A. Statistical Methods for Research Workers. Oliver and Boyd, London, 1925.
- Fogli, Lawrence L., Charles L. Hulin, and Milton R. Blood. Development of first-level behavioral job criteria. Journal of Applied Psychology, Vol. 55, 3-8.
- Gunderson, E. K. E. and P. D. Nelson. Criterion measures for extremely isolated groups. Personnel Psychology, 1966, Vol. 55, 3-8.
- Kavanagh, M. J., A. C. MacKinney, and L. Wolins. Issues in managerial performance: Multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 34-49.
- Larsen, John M. Measurement of the Attitudes of Equivalent, Isolated Groups. Unpublished doctoral dissertation, Purdue University, 1963.
- Latham, G. P., K. N. Wexley, and E. D. Pursell. Training managers to minimize rating errors in the observation

- of behavior. Journal of Applied Psychology, 1975, Vol. 60, 550-555.
- Lawler, E. E. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, Vol. 51, 369-381.
- Levine, J. and J. Butler. Lecture vs. group discussion in changing behavior. Journal of Applied Psychology, 1952, Vol. 36, 29-33.
- McGregor, D. An uneasy look at performance appraisal. Harvard Business Review, 1957, Vol. 35(3), 89-94.
- Mandell, M. M. Supervisory characteristics and ratings: A summary of recent research. Personnel, 1956, Vol. 32, 435-440.
- Nunnally, Jum C. Psychometric Theory. New York, N. Y.: McGraw-Hill Book Company, 1967.
- Nunnally, Jr., Jum C. Introduction to Psychological Measurement. New York, N. Y.: McGraw-Hill Book Company, 1970.
- Porter, L. W., E. E. Lawler, III, and J. R. Hackman. Behavior in Organizations. New York, N. Y.: McGraw-Hill Book Company, 1975.
- Saal, F. E. and F. J. Landy. The mixed standard rating scale: An evaluation. Organizational Behavior and Human Performance, 1977, Vol. 18, 3-18.
- Schneier, C. E. Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 1977, Vol. 62, 541-548.
- Schwab, Donald P., Herbert G. Heneman, III, and Thomas DeCotiis. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, Winter 1975, Vol. 28, 549-562.
- Sisson, B. D. Forced choice: The new army rating. Personnel Psychology, 1948, Vol. 1, 365-381.
- Smith, P. C. and L. M. Kendall. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, Vol. 47, 149-155.

- Stockford, L. and H. W. Bissel. Factors involved in establishing a merit rating-scale. Personnel, 1949, Vol. 26, 94-116.
- Taft, R. The ability to judge people. Psychological Bulletin, 1955, Vol. 52, 1-23.
- Taylor, E. K. and R. Hastman. Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychology, 1956, Vol. 9, 181-206.
- Taylor, E. K. and R. J. Wherry. A study of leniency in two rating systems. Personnel Psychology, 1951, Vol. 4, 25-29.
- Thorndike, E. L. A constant error in psychological ratings. Journal of Applied Psychology, 1920, Vol. 4, 25-29.
- Warmke, D. L. and R. S. Billings. A comparison of training methods for altering the psychometric properties of experimental and administrative performance ratings. Journal of Applied Psychology, 1979, Vol. 64, 124-131.
- Williams, W. E. and D. A. Seiler. Relationship between measures of effort and job performance. Journal of Applied Psychology, 1973, Vol. 57, 49-54.
- Zedeck, S. and H. T. Baker. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 1972, Vol. 7, 457-466.

APPENDIXES

APPENDIX A
PERFORMANCE RATING FORM

PERFORMANCE EVALUATION - CIRCULATION DIVISION

February 6, 1980

Employee _____ Supervisor _____

Job (Circle One) District Manager Relief District Manager Utility District Manager

INSTRUCTIONS: Please use the scale as defined below to evaluate the Performance of the person named above. For each job component listed, two responses are required. FIRST, an evaluation of the person's present performance on the job. SECOND, an indication of your confidence in the evaluation assigned.

For present performance on the job use the appropriate number as defined and described below.

In evaluation procedures we sometimes have good opportunities to observe performance and can have confidence in our evaluation. In other instances the opposite is true. Use the letters as described below to indicate your confidence in your evaluation.

- | | |
|---|--|
| 1 - SUPERIOR ... Performance Superior, usually far exceeds standards and expectations for this district/division. | A. Very high confidence
1. Performance observed very frequently.
2. Good opportunity to assess performance. |
| 2 - EXCELLENT ... High performance, usually moderately exceeds standards and expectations for this district/division. | B. Moderate confidence
1. Performance observed with some frequency.
2. Fair opportunity to assess performance. |
| 3 - GOOD ... Good performance, usually slightly exceeds standards and expectations for this district/division. | C. Low confidence
1. Performance observed infrequently.
2. Poor opportunity to assess performance. |
| 4 - ACCEPTABLE ... Acceptable performance, usually meets standards and expectations for this district/division. | D. No confidence
1. Performance rarely or never observed.
2. Rare or no opportunity to assess performance. |
| 5 - MARGINAL ... Marginal performance, usually meets part of standards and expectations for this district/division. | |
| 6 - POOR ... Poor performance, usually meets few standards and expectations for this district/division. | |
| 7 - UNACCEPTABLE ... Inferior performance fails to meet standards and expectations for this district/division. | |

(Reduced from original size for reproduction)

SERVICE

Circle one number and one letter for each job component listed.

Pre-distribution

1. Checks mailbox for starts, stops and complaints.
2. Has legible run card and route orders.
3. Counts and loads papers at dock.

PerformanceConfidence

1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D

Distribution

1. Distributes papers to drop points, counts, and marks bundles with route and draw number, drops off papers.
2. Protects papers from weather.

1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D

Driving

1. Drives safely with care and excellence - no complaints about driving received.
2. Maintains personal vehicle in excellent mechanical condition; clean and neat.
3. Reports need for maintenance on Company vehicle and keeps clean and neat.
4. Conserves fuel by planning driving and organizing daily activities, drives to conserve fuel.

1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D

Carrier related service

1. Treats carriers fairly and honestly. Works with carriers for prompt, accurate delivery standards; explains to carriers what needs to be done, teaches carriers how to carry and collect. Helps carrier make anticipated profits, goes with carriers on problem collections, teaches carriers how to deal with adults.
2. Sees carriers regularly, checks at carrier pickup points to see that carriers have started, purchases supplies to sell to carriers. Promptly, keeps appointments with carriers and parents on time.

1	2	3	4	5	6	7	A	B	C	D
1	2	3	4	5	6	7	A	B	C	D

<u>SERVICE</u>	<u>Performance</u>	<u>Confidence</u>
3. Helps solve problems, works with carriers on missed deliveries, complaints, and all other problems including carrier's personal problems, if necessary, gets carrier and customer together to solve problems and eliminate hard feelings, works understandingly on broken window cases, etc.	1 2 3 4 5 6 7	A B C D
<u>Personal service on district</u>		
1. Meets standards of Company in coming to work on time, attends department and other meetings on time.	1 2 3 4 5 6 7	A B C D
2. Checks in and out with Supervisor, checks problems with Supervisor and plans work.	1 2 3 4 5 6 7	A B C D
3. Checks and cleans up corners.	1 2 3 4 5 6 7	A B C D
4. Appearance is appropriate for job.	1 2 3 4 5 6 7	A B C D
5. Makes deliveries of shorts, replacement napers, etc., calls in to office three or more times per shift to check problems.	1 2 3 4 5 6 7	A B C D
6. Delivers for no-show carriers and carries down routes, delivers out of the way (off route) singles.	1 2 3 4 5 6 7	A B C D
7. Follows up on complaints and starts of the day and of the day before by contacting customer, rechecks to see that service is satisfactory.	1 2 3 4 5 6 7	A B C D
<u>Route organization</u>		
Arranges routes for convenience and density so that carriers can follow instructions and schedules, checks all routes with carrier to locate all subscribers and non-subscribers, checks route books with draw, including P.I.A. Checks route books for up to date records, sheets in book, etc. Updates Company block list, keeps routes zoned.	1 2 3 4 5 6 7	A B C D
<u>Maintains carrier force</u>		
1. Finding - checks with residents, carriers, dealers, customers for carrier leads, puts out letters, ads, etc. to attract carriers. Keeps a current and always updated list of prospective carriers.	1 2 3 4 5 6 7	A B C D

<u>SERVICE</u>	<u>Performance</u>	<u>Confidence</u>
2. Training - tells new carriers and parents what is expected including substitute before the applicant takes a route, trains carrier on route and checks performance.	1 2 3 4 5 6 7	A B C D
3. Gets bond, contracts and transportation agreements signed from new carriers before carrier starts route.	1 2 3 4 5 6 7	A B C D
<u>Maintain racks</u>		
1. Maintains racks mechanically.	1 2 3 4 5 6 7	A B C D
2. Keeps racks clean.	1 2 3 4 5 6 7	A B C D
3. Loads display papers correctly and keeps rack current.	1 2 3 4 5 6 7	A B C D
4. Removes and counts money and returns, records figures (totals) for each rack.	1 2 3 4 5 6 7	A B C D
OVERALL PERFORMANCE ON SERVICE	1 2 3 4 5 6 7	A B C D

SALESPersonal, direct sales work

1. Establishes new rack and dealer locations.	1 2 3 4 5 6 7	A B C D
2. Makes sales through personal contacts with non-subscribers, sells to isolated customers.	1 2 3 4 5 6 7	A B C D
3. Creates new routes where possible, builds from off routes where feasible.	1 2 3 4 5 6 7	A B C D

Carrier-related, indirect sales work

1. Works for sales through contacting carriers at least weekly, conducts sales training for carriers, accompanies carrier on routes to make list of non-subscribers, informs carriers about quotas, distributes special offer cards.	1 2 3 4 5 6 7	A B C D
2. Promotes contest participation, promotes cold calls, furnishes sales promotion materials.	1 2 3 4 5 6 7	A B C D
OVERALL PERFORMANCE ON SALES	1 2 3 4 5 6 7	A B C D

COLLECTIONS

	<u>Performance</u>							<u>Confidence</u>			
<u>Regular collections</u>											
Delivers P.I.A. printout to carrier promptly, makes regular collections from carriers weekly at pre-determined times and locations, counts and verifies amount paid by carrier in presence of carrier and gives signed bill (receipt).	1	2	3	4	5	6	7	A	B	C	D
<u>Solves collection problems</u>											
Accompanies carrier on route, home to home, to prevent problems, goes to carriers house to prevent problems, encourages carriers to work slow pay collections to not lose money and to make collections on a regular schedule.	1	2	3	4	5	6	7	A	B	C	D
<u>Personal collection work</u>											
Collects down routes, singles and dealers, pursues slow paying customers after consulting with carrier or goes with carrier to slow pay addresses, collects no-pay accounts, collects bad checks written through the district.	1	2	3	4	5	6	7	A	B	C	D
OVERALL PERFORMANCE ON COLLECTIONS	1	2	3	4	5	6	7	A	B	C	D

ACCOUNTING AND RECORDKEEPING

1. Office - prepares various office reports and paper work needed in everyday work, daily check-in sheet, daily worksheets, run book for next day. Completes time sheet.	1	2	3	4	5	6	7	A	B	C	D
2. Relief - leaves instructions for relief manager on days off.	1	2	3	4	5	6	7	A	B	C	D
3. Carrier - checks draw sheet with draw and keeps carriers charged with correct number of papers.	1	2	3	4	5	6	7	A	B	C	D
4. Money - counts money, balances rack and bond sheets, bags coins and turns in rack and bond money to cashier, makes out deposit slips, deposits collected funds, counts money turned in with teller or rechecks later if off hour deposit is used, keeps records of carrier payments.	1	2	3	4	5	6	7	A	B	C	D

ACCOUNTING AND RECORDKEEPINGPerformanceConfidence

5. Carrier bond - keeps careful record of
carrier bond payments.

1 2 3 4 5 6 7

A B C D

OVERALL PERFORMANCE ON ACCOUNTING AND RECORDKEEPING

1 2 3 4 5 6 7

A B C D

EFFORT

In order to evaluate effort we must make different judgments about how hard a person tries, how willing they are to make extra contributions, how attentive to the job are they and many other possible indications of effort. Using the seven point numerical scale and the four letter confidence scale, evaluate the effort made in performance by this person.

PerformanceConfidence

Effort put forth on service

Carrier related service

1 2 3 4 5 6 7

A B C D

Personal service

1 2 3 4 5 6 7

A B C D

Effort put forth on sales related duties

Personal effort

1 2 3 4 5 6 7

A B C D

Effort through carriers

1 2 3 4 5 6 7

A B C D

Effort put forth on accounting and recordkeeping
duties

Office

1 2 3 4 5 6 7

A B C D

Field

1 2 3 4 5 6 7

A B C D

Effort put forth on collection duties and
responsibilities overall

1 2 3 4 5 6 7

A B C D

Effort put forth on carrier force maintenance
duties overall

1 2 3 4 5 6 7

A B C D

Effort put forth on personal service and route
organization duties (i.e., trouble shooting)
overall

1 2 3 4 5 6 7

A B C D

OVERALL PERFORMANCE ON EFFORT

1 2 3 4 5 6 7

A B C D

PREDICTED PERFORMANCE

The next task for you as an evaluator is to make an estimate of the probable performance of this person on future assignments. Use your knowledge of the individual and your experience with that individual to estimate how well that person will perform the task components described.

Use the following scale numbers to indicate your prediction. Please do not look back at your evaluation of current performance.

- 1 - SUPERIOR ... Learns extremely quickly, expected to far exceed standards and/or expectations promptly.
- 2 - EXCELLENT ... Learns very quickly, expected to moderately exceed standards and/or expectations within a short time.
- 3 - GOOD ... Learns well, expected to slightly exceed standards and/or expectations.
- 4 - ACCEPTABLE ... Learns moderately well, expected to meet standards and/or expectations.
- 5 - MARGINAL ... Learns somewhat slowly, expected to meet part of standards and/or expectations.
- 6 - POOR ... Slow to learn, expected to meet few standards and/or expectations.
- 7 - INFERIOR ... Not at all suitable for this work, cannot meet standards or expectations.

SERVICE

Circle one number for each job component listed.

Pre-distributionPredicted Performance

1. Checks mailbox for starts, stops and complaints.
2. Has legible run card and route orders.
3. Counts and loads papers at dock.

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Distribution

1. Distributes papers to drop points, counts, and marks bundles with route and draw number, drops off papers.
2. Protects papers from weather.

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Driving

1. Drives safely with care and excellence - no complaints about driving received.
2. Maintains personal vehicle in excellent mechanical condition; clean and neat.
3. Reports need for maintenance on Company vehicle and keeps clean and neat.
4. Conserves fuel by planning driving and organizing daily activities, drives to conserve fuel.

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Carrier related service

1. Treats carriers fairly and honestly. Works with carriers for prompt, accurate delivery standards; explains to carriers what needs to be done, teaches carriers how to carry and collect. Helps carrier make anticipated profits, goes with carriers on problem collections, teaches carriers how to deal with adults.
2. Sees carriers regularly, checks at carrier pickup points to see that carriers have started, purchases supplies to sell to carriers. Promptly, keeps appointments with carriers and parents on time.

1 2 3 4 5 6 7

1 2 3 4 5 6 7

SERVICEPredicted Performance

3. Helps solve problems, works with carriers on missed deliveries, complaints, and all other problems including carrier's personal problems, if necessary, gets carrier and customer together to solve problems and eliminate hard feelings, works understandingly on broken window cases, etc.

1 2 3 4 5 6 7

Personal service on district

1. Meets standards of Company in coming to work on time, attends department and other meetings on time.
2. Checks in and out with Supervisor, checks problems with Supervisor and plans work.
3. Checks and cleans up corners.
4. Appearance is appropriate for job.
5. Makes deliveries of shorts, replacement papers, etc., calls in to office three or more times per shift to check problems.
6. Delivers for no-show carriers and carries down routes, delivers out of the way (off route) singles.
7. Follows up on complaints and starts of the day and of the day before by contacting customer, rechecks to see that service is satisfactory.

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

1 2 3 4 5 6 7

Route organization

Arranges routes for convenience and density so that carriers can follow instructions and schedules, checks all routes with carrier to locate all subscribers and non-subscribers, checks route books with draw, including P.I.A. Checks route books for up to date records, sheets in book, etc. Updates Company block list, keeps routes zoned.

1 2 3 4 5 6 7

Maintains carrier force

1. Finding - checks with residents, carriers, dealers, customers for carrier leads, puts out letters, ads, etc. to attract carriers. Keeps a current and always updated list of prospective carriers.

1 2 3 4 5 6 7

SERVICEPredicted Performance

2. Training - tells new carriers and parents what is expected including substitute before the applicant takes a route, trains carrier on route and checks performance.

1 2 3 4 5 6 7

3. Gets bond, contracts and transportation agreements signed from new carriers before carrier starts route.

1 2 3 4 5 6 7

Maintain racks

1. Maintains racks mechanically.

1 2 3 4 5 6 7

2. Keeps racks clean.

1 2 3 4 5 6 7

3. Loads display papers correctly and keeps rack current.

1 2 3 4 5 6 7

4. Removes and counts money and returns, records figures (totals) for each rack.

1 2 3 4 5 6 7

OVERALL PERFORMANCE ON SERVICE

1 2 3 4 5 6 7

SALESPersonal, direct sales work

1. Establishes new rack and dealer locations.

1 2 3 4 5 6 7

2. Makes sales through personal contacts with non-subscribers, sells to isolated customers.

1 2 3 4 5 6 7

3. Creates new routes where possible, builds from off routes where feasible.

1 2 3 4 5 6 7

Carrier-related, indirect sales work

1. Works for sales through contacting carriers at least weekly, conducts sales training for carriers, accompanies carrier on routes to make list of non-subscribers, informs carriers about quotas, distributes special offer cards.

1 2 3 4 5 6 7

2. Promotes contest participation, promotes cold calls, furnishes sales promotion materials.

1 2 3 4 5 6 7

OVERALL PERFORMANCE ON SALES

1 2 3 4 5 6 7

COLLECTIONSPredicted PerformanceRegular collections

Delivers P.I.A. printout to carrier promptly, makes regular collections from carriers weekly at pre-determined times and locations, counts and verifies amount paid by carrier in presence of carrier and gives signed bill (receipt).

1 2 3 4 5 6 7

Solves collection problems

Accompanies carrier on route, home to home, to prevent problems, goes to carriers house to prevent problems, encourages carriers to work slow pay collections to not lose money and to make collections on a regular schedule.

1 2 3 4 5 6 7

Personal collection work

Collects down routes, singles and dealers, pursues slow paying customers after consulting with carrier or goes with carrier to slow pay addresses, collects no-pay accounts, collects bad checks written through the district.

1 2 3 4 5 6 7

OVERALL PERFORMANCE ON COLLECTIONS

1 2 3 4 5 6 7

ACCOUNTING AND RECORDKEEPING

1. Office - prepares various office reports and paper work needed in everyday work, daily check-in sheet, daily worksheets, run book for next day. Completes time sheet.

1 2 3 4 5 6 7

2. Relief - leaves instructions for relief manager on days off.

1 2 3 4 5 6 7

3. Carrier - checks draw sheet with draw and keeps carriers charged with correct number of papers.

1 2 3 4 5 6 7

4. Money - counts money, balances rack and bond sheets, bags coins and turns in rack and bond money to cashier, makes out deposit slips, deposits collected funds, counts money turned in with teller or rechecks later if off hour deposit is used, keeps records of carrier payments.

1 2 3 4 5 6 7

ACCOUNTING AND RECORDKEEPINGPredicted Performance

5. Carrier bond - keeps careful record of
carrier bond payments.

1 2 3 4 5 6 7

OVERALL PERFORMANCE ON ACCOUNTING AND RECORDKEEPING

1 2 3 4 5 6 7

APPENDIX B
MEANS AND STANDARD DEVIATIONS

TABLE 9
ITEM MEANS AND STANDARD DEVIATIONS FOR
DEPARTMENT HEAD PERFORMANCE RATINGS

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
<u>Service</u>				
Predistribution	1	2.06	.95	122
	2	2.32	.99	119
	3	2.14	.96	123
Distribution	1	2.08	1.06	123
	2	2.02	1.01	123
Driving	1	1.97	1.13	123
	2	2.04	1.06	94
	3	2.51	.85	35
	4	2.72	.85	123
Carrier related service	1	2.83	1.18	88
	2	2.67	1.04	90
	3	2.86	1.08	90
Personal service on district	1	1.93	1.15	123
	2	2.20	1.07	123
	3	2.48	.88	116
	4	2.11	1.30	123
	5	2.24	1.03	51
	6	2.26	1.08	113
	7	2.89	1.23	114
Route organization		2.80	1.18	76
Maintains carrier force	1	3.00	1.30	84
	2	2.92	1.21	88
	3	2.58	1.23	86
Maintain racks	1	2.41	.83	73
	2	2.99	.79	73
	3	1.93	.88	102
	4	1.82	.95	73
<u>Overall Performance on Service</u>		2.60	.83	120
<u>Sales</u>				
Personal, Direct sales work	1	3.58	.94	91
	2	3.82	1.13	114
	3	3.65	1.07	71
Carrier-related, indirect sales work	1	3.65	1.04	83
	2	3.43	1.20	88
<u>Overall Performance on Sales</u>		3.76	1.06	119

TABLE 9 (Cont'd)

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
<u>Collections</u>				
Regular collections		2.47	.98	85
Solves collection problems	1	2.64	1.09	84
Personal collection work	1	2.81	1.04	111
<u>Overall Performance on Collections</u>				
		2.77	.93	111
<u>Accounting and Recordkeeping</u>				
	1	2.12	1.06	123
	2	2.23	1.06	103
	3	2.21	1.09	116
	4	2.00	1.04	99
	5	1.98	1.05	83
<u>Overall Performance on Accounting and Recordkeeping</u>				
		2.27	.98	122
<u>Effort</u>				
Effort put forth on service	1	2.97	1.08	90
	2	2.73	1.09	122
Effort put forth on sales	1	3.57	1.26	118
related duties	2	3.63	1.18	88
Effort put forth on accounting	1	2.21	1.06	122
and recordkeeping duties	2	2.58	1.00	117
Effort put forth on collection		2.39	1.12	115
Effort put forth on carrier		2.74	1.34	85
force maintenance				
Effort put forth on personal		2.86	1.08	98
service and route organiza-				
tion				
<u>Overall Performance on Effort</u>				
		2.83	.90	121

TABLE 10

ITEM MEANS AND STANDARD DEVIATIONS FOR FIRST
LINE SUPERVISOR PERFORMANCE RATINGS

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
<u>Service</u>				
Predistribution	1	2.15	1.01	122
	2	2.40	1.09	118
	3	2.20	1.10	123
Distribution	1	2.33	.95	117
	2	2.07	.94	117
Driving	1	2.31	1.16	123
	2	2.30	1.19	97
	3	2.55	.87	44
	4	2.72	1.04	117
Carrier related service	1	2.93	1.04	92
	2	2.79	.90	97
	3	2.99	.93	94
Personal service on district	1	2.27	1.29	123
	2	2.33	1.07	123
	3	2.56	.98	115
	4	2.63	1.15	123
	5	2.51	1.11	67
	6	2.32	1.10	112
	7	3.00	1.20	114
Route organization		3.02	.97	82
Maintains carrier force	1	3.10	1.18	87
	2	2.97	1.09	89
	3	2.72	1.12	87
Maintain racks	1	2.46	.91	87
	2	3.49	1.16	88
	3	2.30	1.15	98
	4	2.39	1.17	95
<u>Overall Performance on Service</u>		2.79	.83	119
<u>Sales</u>				
Personal, direct sales work	1	3.63	1.02	107
	2	3.57	1.17	115
	3	3.38	1.17	79
Carrier-related, indirect sales work	1	3.38	1.00	91
	2	3.24	.94	95
<u>Overall Performance on Sales</u>		3.56	.98	115
<u>Collections</u>				
Regular collections		2.42	1.09	89

TABLE 10 (Cont'd)

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
Solves collection problems	1	2.89	.97	87
Personal collection work	1	2.84	1.04	108
<u>Overall Performance on Collections</u>		2.75	.93	105
<u>Accounting and Recordkeeping</u>	1	2.52	1.29	122
	2	2.27	1.17	107
	3	2.44	1.20	111
	4	2.21	1.03	105
	5	2.10	1.01	90
<u>Overall Performance on Accounting and Recordkeeping</u>		2.48	1.06	117
<u>Effort</u>				
Effort put forth on service	1	2.98	.92	101
	2	2.78	.97	122
Effort put forth on sales related duties	1	3.57	1.06	120
	2	3.50	1.00	100
Effort put forth on account- ing and recordkeeping duties	1	2.61	1.14	123
	2	2.71	1.05	119
Effort put forth on collection		2.78	.97	117
Effort put forth on carrier force maintenance		2.98	1.27	93
Effort put forth on personal service and route organiza- tion		3.07	1.09	98
<u>Overall Performance on Effort</u>		3.03	.89	122

TABLE 11

ITEM MEANS AND STANDARD DEVIATIONS FOR DEPARTMENT
HEAD PREDICTED PERFORMANCE RATINGS

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
<u>Service</u>				
Predistribution	1	1.73	.85	123
	2	1.93	.92	122
	3	1.77	.89	123
Distribution	1	1.80	.93	123
	2	1.74	.88	123
Driving	1	1.79	.92	123
	2	1.95	1.05	102
	3	2.00	.74	66
	4	2.50	.96	123
Carrier related service	1	2.60	1.09	101
	2	2.60	1.02	101
	3	2.75	1.08	102
Personal service on district	1	1.76	1.10	123
	2	2.06	.99	123
	3	2.21	.95	122
	4	1.89	1.10	123
	5	2.20	.85	66
	6	2.17	.89	118
	7	2.68	1.03	121
Route organization		2.78	1.07	93
Maintains carrier force	1	2.89	1.22	100
	2	2.88	1.15	100
	3	2.40	1.14	100
Maintain racks	1	2.25	.91	101
	2	2.60	.84	101
	3	1.76	.84	119
	4	1.81	.87	98
<u>Overall Performance on Service</u>		2.39	.90	122
<u>Sales</u>				
Personal, direct sales work	1	3.14	.99	101
	2	3.29	1.24	119
	3	3.32	1.06	93
Carrier-related, indirect sales work	1	3.20	1.17	100
	2	3.10	1.18	101
<u>Overall Performance on Sales</u>		3.30	1.19	120
<u>Collections</u>				
Regular collections		2.45	1.01	100

TABLE 11 (Cont'd)

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
Solves collection problems		2.67	1.08	100
Personal collection work		2.61	1.06	113
<u>Overall Performance on Collections</u>		2.60	.99	114
Accounting and Recordkeeping	1	1.88	.95	123
	2	2.07	1.01	107
	3	1.94	.98	122
	4	1.90	1.01	106
	5	1.91	.98	100
<u>Overall Performance on Accounting and Recordkeeping</u>		2.05	.98	123

TABLE 12

ITEM MEANS AND STANDARD DEVIATIONS FOR FIRST LINE
SUPERVISOR PREDICTED PERFORMANCE RATINGS

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
<u>Service</u>				
Predistribution	1	1.96	.95	123
	2	2.21	1.01	122
	3	2.05	1.05	123
Distribution	1	2.19	.98	123
	2	2.02	.93	122
Driving	1	2.17	1.08	123
	2	2.25	1.07	117
	3	2.25	.84	85
	4	2.53	1.00	116
Carrier related service	1	2.82	1.02	114
	2	2.75	1.00	114
	3	2.86	1.01	116
Personal service on district	1	2.12	1.17	123
	2	2.07	.98	123
	3	2.24	.88	122
	4	2.31	1.00	123
	5	2.32	.96	106
	6	2.15	.99	121
	7	2.77	1.00	121
Route organization		3.01	1.00	111
Maintains carrier force	1	2.96	1.08	112
	2	2.88	1.00	112
	3	1.92	.91	110
	4	2.13	1.03	110
<u>Overall Performance on Service</u>		2.57	.90	121
<u>Sales</u>				
Personal, direct sales work	1	3.15	.96	117
	2	3.28	1.11	120
	3	3.27	1.00	113
Carrier-related, indirect sales work	1	3.10	.96	113
	2	3.09	1.01	115
<u>Overall Performance on Sales</u>		3.21	.99	120
<u>Collections</u>				
Regular collections		2.50	1.03	114
Solves collection problems		2.75	.99	114
Personal collection work		2.64	.96	119
<u>Overall Performance on Collections</u>		2.65	.95	118

TABLE 12 (Cont'd)

Dimension/Subdivision	Item Number	Mean	Standard Deviation	N
<u>Accounting and Recordkeeping</u>	1	2.19	1.08	123
	2	2.19	1.09	118
	3	2.13	1.02	119
	4	2.10	.94	117
	5	2.12	1.01	115
<u>Overall Performance on Accounting and Recordkeeping</u>		2.30	1.05	121

APPENDIX C
STATISTICAL DATA

TABLE 13

ITEM CORRELATIONS FOR PERFORMANCE RATINGS WITH
CONFIDENCE RATINGS FOR FIRST LINE SUPERVISOR

Dimension/Subdivision	Item Number	N	r
<u>Service</u>			
Predistribution	1	121	.29**
	2	117	.20*
	3	119	.16
Distribution	1	115	.27*
	2	115	.14
Driving	1	121	.08
	2	95	.16
	3	42	.23
	4	116	-.05
Carrier related service	1	92	-.10
	2	96	.16
	3	93	.06
Personal service on district	1	121	.00
	2	121	.18*
	3	113	.40**
	4	120	.16
	5	65	.38*
	6	110	.37**
	7	112	.11
Route organization	1	82	.10
Maintains carrier force	1	86	.12
	2	88	.06
	3	85	.21
Maintain racks	1	85	.24*
	2	87	-.07
	3	95	.31*
	4	93	.16
<u>Overall Performance on Service</u>		111	-.01
<u>Sales</u>			
Personal, direct sales work	1	105	-.17
	2	110	.14
	3	77	-.11
Carrier-related, indirect sales work	1	86	.11
	2	89	.19
<u>Overall Performance on Sales</u>		105	.04
<u>Collections</u>			
Regular collections	1	89	-.07

TABLE 13 (Cont'd)

Dimension/Subdivision	Item Number	N	r
Solves collection problems	1	87	.02
Personal collection work	1	108	-.02
<u>Overall Performance on Collections</u>		90	.02
<u>Accounting and Recordkeeping</u>	1	119	.18
	2	105	.14
	3	107	.21*
	4	103	.34**
	5	87	.36**
<u>Overall Performance on Accounting and Recordkeeping</u>		107	.12
<u>Effort</u>			
Effort put forth on service	1	100	.16
	2	121	.21*
Effort put forth on sales related duties	1	117	.17
	2	99	.00
Effort put forth on accounting and recordkeeping duties	1	121	.19*
	2	117	.23*
Effort put forth on collections	1	113	.08
Effort put forth on carrier force maintenance	1	93	.11
Effort put forth on personal service and route organization	1	96	.23*
<u>Overall Performance on Effort</u>		115	-.03

*p < .05

**p < .001

TABLE 14

ITEM CORRELATIONS FOR PERFORMANCE RATINGS WITH
CONFIDENCE RATINGS FOR DEPARTMENT HEAD

Dimension/Subdivision	Number	N	r
<u>Service</u>			
Predistribution	1	113	.21*
	2	113	.31**
	3	113	.26*
Distribution	1	117	.48**
	2	117	.36**
Driving	1	116	.09
	2	98	.25*
	3	30	.15
	4	117	-.11
Carrier related service	1	88	.13
	2	90	.13
	3	90	.14
Personal service on district	1	117	.12
	2	117	.05
	3	116	.32**
	4	117	.02
	5	51	.62**
	6	107	.32**
	7	114	.18*
Route organization	1	76	-.08
Maintains carrier force	1	84	-.11
	2	87	.22*
	3	85	.15
Maintain racks	1	64	.05
	2	64	-.27*
	3	94	.39**
	4	65	.15
<u>Overall Performance on Service</u>		109	.08
<u>Sales</u>			
Personal, direct sales work	1	85	-.15
	2	108	.06
	3	71	-.16
Carrier-related, indirect sales work	1	83	.06
	2	88	.45**
<u>Overall Performance on Sales</u>		113	.21*
<u>Collections</u>			
Regular collections	1	85	.14
Solves collection problems	1	84	.28*

TABLE 14 (Cont'd)

Dimension/Subdivision	Number	N	r
Personal collection work	1	104	.16
<u>Overall Performance on Collections</u>		105	.18
<u>Accounting and Recordkeeping</u>	1	117	.15
	2	98	.29*
	3	109	.13
	4	93	.14
	5	78	.26*
<u>Overall Performance on Accounting and Recordkeeping</u>		112	.17
<u>Effort</u>			
Effort put forth on service	1	89	.21*
	2	115	.17
Effort put forth on sales	1	110	.14
related duties	2	86	.02
Effort put forth on accounting and	1	115	.11
and recordkeeping duties	2	110	.11
Effort put forth on collections	1	108	.04
Effort put forth on carrier	1	83	.18
force maintenance			
Effort put forth on personal	1	91	.20
service and route organization			
<u>Overall Performance on Effort</u>		114	.04

*p < .05

**p < .001

TABLE 15

ITEM CORRELATIONS FOR PERFORMANCE RATINGS WITH
CONFIDENCE RATINGS FOR BOTH SUPERVISORS
COMBINED

Dimension/Subdivision	Item Number	N	r
<u>Service</u>			
Predistribution	1	234	.16*
	2	230	.17*
	3	232	.15*
Distribution	1	232	.21**
	2	232	.17*
Driving	1	237	-.01
	2	185	.08
	3	72	.14
	4	233	-.07
Carrier related service	1	180	.00
	2	186	.09
	3	183	.06
Personal service on district	1	238	.03
	2	238	.04
	3	229	.23**
	4	237	.05
	5	117	.43**
	6	217	.26**
	7	226	.10
Route organization	1	158	-.06
Maintains carrier force	1	170	-.03
	2	175	.11
	3	170	.14
Maintain racks	1	149	.10
	2	151	-.25*
	3	189	.15*
	4	158	.00
<u>Overall Performance on Service</u>		220	-.02
<u>Sales</u>			
Personal, direct sales work	1	190	-.14
	2	218	.15*
	3	148	-.05
Carrier-related, indirect sales work	2	117	.34**
<u>Overall Performance on Sales</u>		218	.18*
<u>Collections</u>			
Regular collections	1	174	.05

TABLE 15 (Cont'd)

Dimension/Subdivision	Item Number	N	r
Solves collection problems	1	171	.08
Personal collection work	1	212	.06
<u>Overall Performance on Collections</u>		195	.09
<u>Accounting and Recordkeeping</u>	1	236	.07
	2	203	.18*
	3	216	.08
	4	196	.15*
	5	165	.21*
<u>Overall Performance on Accounting and Recordkeeping</u>		219	.04
<u>Effort</u>			
Effort put forth on service	1	189	.17*
	2	236	.16*
Effort put forth on sales related duties	1	227	.15*
	2	185	.03
Effort put forth on accounting and recordkeeping duties	1	236	.04
	2	227	.07
Effort put forth on collections	1	221	-.02
Effort put forth on carrier force maintenance	1	176	.09
Effort put forth on personal service and route organization	1	187	.16*
<u>Overall Performance on Effort</u>		229	-.04

*p < .05

**p < .001

VITA

Marvin Lee Stewart was born March 12, 1947, in Memphis, Tennessee. He attended grade school there and graduated from Whitehaven High School in 1965. In the fall of that year he enrolled in The University of Tennessee, Knoxville. In 1968 he interrupted his education to join the U. S. Navy and served as an aviation radar electronics' technician for four years. He returned to The University of Tennessee, Knoxville, and received a B.S. degree in Industrial and Personnel Management in 1974. While attending the university he was a member of Acacia social fraternity. Geneva Abell and Lee met at the university and were married in 1975. They had their first child, Alexander, in August of 1981. Lee was a candidate for a M.S. degree in Industrial and Organizational Psychology in June 1982.

In his work experience he has performed such jobs as electronics technician, computer programmer, and assistant manager for a music shop. He assisted in the development of an employee selection program at the Tennessee Valley Authority. In the present study he has worked in the area of performance evaluation.