A STUDY OF APPRAISAL METHODOLOGY: THE EFFECT OF THE COORDINATOR IN APPRAISAL

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
Norman Frisbey
1958

THESIS

This is to certify that the

thesis entitled

A Study of Appraisal Methodology:

The Effect of the Coordinator in Appraisal

presented by

Norman Frisbey

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Psychology

Carl T. Frost
Major professor

Date February 19, 1958

O-169

A STUDY OF APPRAISAL METHODOLOGY:
THE EFFECT OF THE COORDINATOR IN APPRAISAL

By

NORMAN FRISBEY

A THESIS

Submitted to the School for Advanced Graduate Studies of
Michigan State University of Agriculture and
Applied Science in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

1958

6-19-59
G 5845

ACKNOWLEDGMENTS

The writer wishes to express his sincere thanks to Dr. Carl F. Frost, major professor, for the guidance and valuable help he provided during the program of study and the course of this research.

Each of the author's Guidance Committee members, Drs. D. M. Johnson, M. Ray Denny, H. C. Smith, and E. H. Jacobson, have made suggestions and contributions for which he is grateful.

The writer appreciates the opportunity provided by Chrysler Corporation to conduct this investigation within its operation.

Special thanks are due Mr. Wayne E. Grimm, Director of Management Development, Chrysler Central Personnel Staff, the person responsible for providing the opportunity to do the research and for making the data available for use in this study. In addition, Mr. Grimm gave direct assistance by providing the experimenter with the personal time required to do the work.

Dr. Edwin F. Harris, Research Assistant, Chrysler Central Personnel Staff, showed a personal interest in the problem and gave advice on technical and operational matters. The writer is grateful for the assistant provided by Dr. Harris to help collect the data.

The writer is indebted to Dr. John Versace, Chrysler Engineering Division, for his advice on the statistical problems.

Thanks go to the Personnel Staff at the Jet Engine Plant for their cooperation and help during the collection of the data.

It is impossible to fully express my gratitude and thanks to my wife, Ardeth, for typing this manuscript and for her encouragement and help throughout the period of graduate school and during the time of this research.

Norman Frisbey
candidate for the degree of
Doctor of Philosophy

Final Examination: February 19, 1958, 2:00 P.M.

Dissertation: A Study of Appraisal Methodology:
The Effect of the Coordinator in Appraisal

Outline of Studies:

Degree	Major	Minor	Granted
B.S.	Mathematics	Psychology	March, 1950
M.A.	Psychology	Statistics	August, 1952
Ph.D.	Psychology		March, 1958

Biographical Items:

Born, June 8, 1920, Ferndale, Michigan

Undergraduate Studies, Michigan State College, 1946-1950

Graduate Studies, Michigan State University, 1950-1958

Experience:

Part-time Instructor, Michigan State College
Spring and Fall - 1954

Research Specialist, Chrysler Corporation
Central Personnel Staff - 1955-1958

Membership:

American Psychological Association
Midwestern Psychological Association
Michigan Psychological Association

A STUDY OF APPRAISAL METHODOLOGY:

THE EFFECT OF THE COORDINATOR IN APPRAISAL

By

NORMAN FRISBEY

AN ABSTRACT

Submitted to the School for Advanced Graduate Studies of
Michigan State University of Agriculture and
Applied Science in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology

1958

Approved Carl T. Frost

Sixty-four first line supervisors, in an industrial plant, each rated or appraised the job performance of three subordinates. Two methods of administration, two coordinators, and two locations were incorporated into a factorial design. Method of administration, which involved comparing coordinated with non-coordinated appraisals, was the variable of primary interest. The coordinator was a personnel staff person who conducted the appraisals by questioning the appraiser and recording the responses in a modified Field Review type interview.

The objectives of appraisal were: (1) evaluation of present performance, and (2) planning for individual improvement. It was predicted that the coordinated appraisals would be superior to the non-coordinated in meeting these aims.

The relative merits of the methods as a system of evaluation were inferred by comparing the treatment groups on secondary criteria of a rating method. It was predicted that the ratings of the coordinated appraisals would be improved by increased discrimination, reduced leniency, reduced halo, increased coverage, and increased comparability between ratings. The findings did not support the above predictions. No significant differences were found between the methods, coordinators or locations on these factors. It was concluded that the coordinator did not improve the effectiveness of the appraisal as a rating instrument.

Immediate criteria used to evaluate the appraisal as a development instrument were established from an examination of its function in the development procedure. Subsequent to the appraisal session the information recorded on the form was to be used for (1) a review by the appraiser's supervisor, (2) a performance interview with the employee, (3) training and development of the subordinate, and (4) follow-up on action by the coordinator. The areas of the form important to the above steps were the supporting facts and performance summary which contained sections for a summary of individual performance strengths, development needs and development plans.

It was predicted that the coordinated appraisals would be superior to the non-coordinated in terms of the quantity and quality of responses recorded in these sections. The findings did support these predictions. The supporting facts of the coordinated appraisals contained a greater amount of information and were more descriptive of specific performance than those of the non-coordinated group. The non-coordinated group of appraisals contained a greater number of appraisals with Sections 1, 2, and 3 of the performance summary omitted. The performance summaries of the coordinated appraisals contained a larger number of performance strengths, development needs, and methods of handling development needs. The development needs for the coordinated group were more frequently related to the supporting facts. The coordinated appraisals more frequently

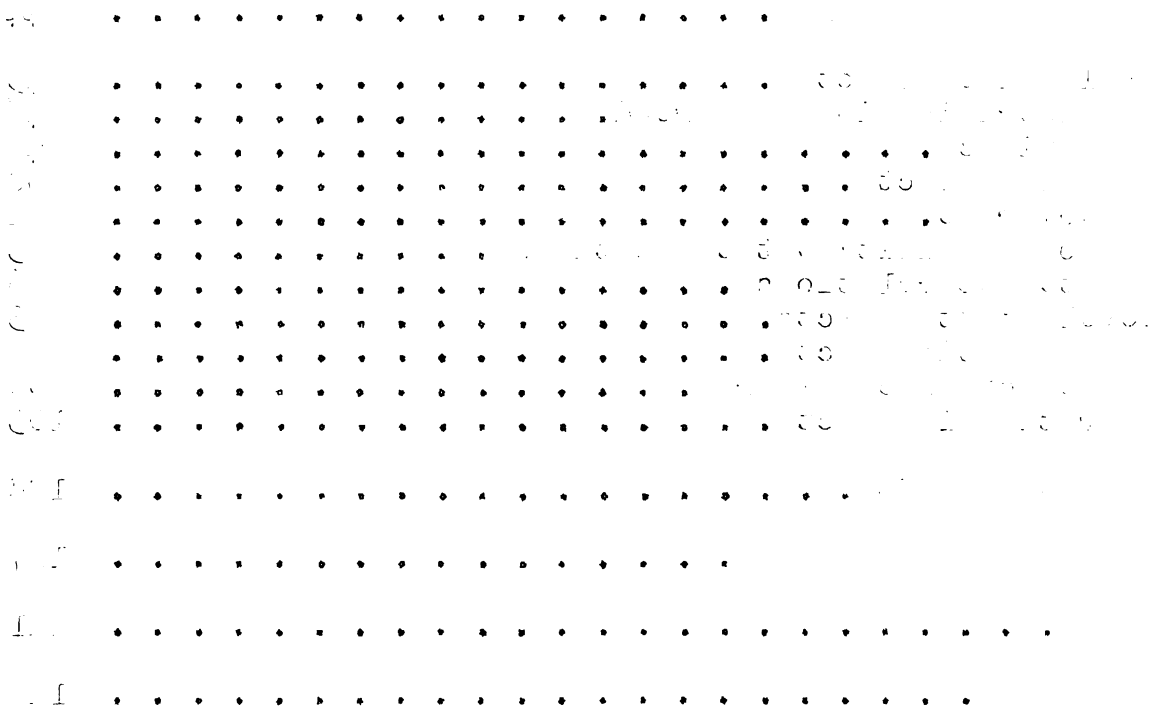
placed responsibility for development action on the supervisor and more frequently mentioned on-the-job coaching as a method of development than the non-coordinated appraisals.

It was concluded that the coordinator did effect an improvement on the appraisals as an instrument for development in terms of the procedure outlined and the criteria used for evaluation. In fact, it can be said that the coordinator plays an essential role and makes a significant contribution to the development procedure.

[illegible]

TABLE OF CONTENTS

TABLE OF CONTENTS Continued



LIST OF TABLES

LIST OF TABLES Continued

LIST OF TABLES Continued

LIST OF FIGURES

INTRODUCTION

General Background

Rating methods, which are today among the recognized means of measuring conduct or behavior are almost as old as any experimental methods in psychology. The beginnings of rating methods can be traced back to the work of Fechner and others in psychophysics. In fact, the method of paired comparisons, a much used technique today, was developed from Fechner's method of impression in 1894. Symonds (35) stated that the first rating scale, in its modern sense, was that published by Galton in 1883. This scale even today may be considered a model of its kind.

One significant development of rating was the man-to-man principle developed by Walter Dill Scott of the Carnegie Institute of Technology during World War I. The Army Rating Scale using this principle was first framed by Scott in May, 1917. Soon after this, in 1920, the true graphic scale was developed by the Scott Company Laboratory. A more complete historical treatment of rating scale developments can be found in Symonds (35).

Rating methods have been widely used for a number of years both for purposes of investigation and research and in practical ways to evaluate personnel in industry, school, armed forces, and civil service. A bibliography of merit

rating by Mahler (23) covering the years 1926-1946 contains approximately 500 references. The recent work of psychologists and personnel research persons since World War II has been characterized by an increasing number of articles and studies concerned with the problems of rating. Although much has been written and many innovations suggested, some of the basic problems which faced the early researchers still remain essentially unsolved.

For thorough coverage of the various types of rating techniques, purposes of rating, pitfalls, advantages and disadvantages of rating, the reader is referred to industrial psychology texts such as Bellows (2) and Tiffin (40) or general tests such as Poffenberger (26) and Symonds (35). Two practical handbooks of rating are Dooher and Marquis (6) and Smyth and Murphy (29). In addition, information can be found in the general management type periodical such as Personnel and Personnel Journal. Only the general background material pertinent to the method under investigation has been presented by the investigator.

Chrysler Appraisal System

It was the purpose of this study to investigate experimentally a rating or appraisal method which was devised to meet some of the short-comings common to rating systems. This particular method is the appraisal system used by the Chrysler Corporation as part of its Management Development Program.

The commonly stated purposes of an appraisal or merit rating program are the following:

1. Promotion or transfer. Management is able to identify persons capable of greater responsibility or with skills needed for other positions.
2. Employee improvement. The strong and weak points of employees are identified so that both management and the employee can direct their efforts toward the development of skills for increased efficiency.
3. Research. The identification of better and poorer groups of employees serves as criteria for the validation of selection procedures.

The aims of the Chrysler Management Development Program are consistent with those above and broadly stated are to improve managerial competency and to build a reserve of well-trained management personnel. This is accomplished through an evaluation of present performance and planned development action on an individual basis. The appraisal and development of management persons includes the following steps:

1. Performance appraisal of the management person by his immediate superior
2. Appraisal review by the appraiser's immediate superior
3. Performance interview with the subordinate appraised
4. Training and development of the subordinate.

The focus of attention for this study was directed toward the performance appraisal, and more specifically, the appraisal session itself. The action indicated by the steps following the appraisal or the utilization of the appraisal information is dependent upon the appraisal session.

The management appraisal instrument is a five point graphic scale containing 35 to 40 items. The items vary somewhat in wording from plant to plant, but the content is essentially the same. The items represent managerial responsibilities which are characteristic of a level of supervision rather than specific positions. These items are arranged under four general areas of responsibility:

1. Planning and directing operations
2. Maintaining a working force
3. Controlling costs
4. Organizational relationships.

Prior to the assignment of a rating evidence of performance pertinent to the item is recorded in a space provided and labeled "Supporting Facts". After rating the entire group of items, the rater assigns an overall performance rating in terms of the same five-category scale as used for the items. This overall rating is judgmental and does not represent any quantitative combination of item ratings.

The next step in the appraisal process is a "Performance Summary". Here the supporting facts for some of the higher ratings are listed as "Strengths" of the ratee. Likewise,

some of the evidence for lower ratings is listed and called "Development Needs". A third section in the summary is titled "Development Plans". In the latter section development plans or recommended courses of action are suggested to help meet the needs previously mentioned.

The distinctive aspect of the appraisal session is that a personnel staff person assists the rater or supervisor in the appraisal of each of his subordinates. This staff man is called a Management Development Coordinator.

The function of the coordinator in the appraisal situation itself is crucial to subsequent phases of activity. With the appraisal instrument as a guide, the coordinator interviews the appraiser to assist him in the evaluation, performance summary and development planning for each subordinate. The role of the coordinator in the appraisal session is outlined briefly as follows:

1. To conduct the appraisal interview and record the ratings and supporting facts
2. To interpret the items and to clarify as to specific application
3. To ask pertinent probe questions which assist in complete coverage
4. To assure that the ratings and facts represent typical rather than infrequent behavior
5. To encourage a spread in the ratings by distinguishing between degrees of performance

6. To encourage that the items be considered as distinct from one another
7. To assist in a summary of performance
8. To guide in determination of an overall rating and potential rating
9. To assist in outlining a sound and attainable individual development plan.

The coordinator is a person thoroughly acquainted with the appraisal system. His participation in the appraisal session is expected to increase:

1. The rater's understanding of the procedures and materials
2. The objectivity or clarity of judgments
3. The coverage of pertinent information
4. The uniformity of application from rater to rater.

There are about sixty persons acting as full-time coordinators in Chrysler where the management force totals approximately twelve thousand.

It is difficult to determine how extensively the coordinator concept is being used by other companies. The usual surveys of rating methods such as those by Spicer (30) and Benjamin (3) do not give information of this type. Surveys do not usually indicate procedures unique to one or two respondents in the sample.

The investigator has personal knowledge that a large manufacturer of electronic products uses coordinators in its

management development program. This company's manual stated that the coordinator acts as the secretary "recording the remarks which the supervisor wishes to make and answering any questions which he may have regarding the use of the form". He also assists on the summaries and conducts the reviews with the personnel supervisor and head of the activity involved. However, it is suggested that after experience, the supervisor make the appraisals alone and that the coordinator go over the forms with him subsequent to rating.

General Mills, Inc. follows this practice. Balch stated,

> We use a trained interviewer to question each appraiser individually concerning a man and to record the appraiser's opinions. The interviewers frequently are personnel men, but some of the most successful have been operating men who were trained and used for short periods on this work. (1, p. 12)

The Dartnell Corporation materials developed by Robert N. McMurry and Company, a management consulting firm, suggested a Patterned Merit Review. This is conducted by a personnel department or home office representative. However, their manual stated, "The interviewer should never interject his own views. He is not evaluating the employee; he is merely providing the supervisor with a frame of reference" (21, p. 4).

The above programs appear to use a coordinator in the appraisal session in a way similar to that of Chrysler. All obtain supporting information for each item rated, summarize strengths and weaknesses and obtain an overall rating.

Field Review Method

Historically the notion of a coordinated appraisal appears to be derived from the Field Review Method of Employee Valuation and Internal Placement. Wadsworth (41), in a series of six articles, presented the rationale and procedure for use of this method of evaluation. He indicated that this procedure was first published in manual form by the Army Service Forces (ASF M_{213}, May 1945) from his pen. Wadsworth further stated the method itself was an outgrowth of extended employee and internal placement research carried on initially in Southern California Gas Company and Southern Counties Gas Company of California in the 1930's.

As outlined by Wadsworth the Field Review Method "is essentially a program of planned supervisor contacts undertaken to make the employee evaluation and placement job effective, and includes everything that ordinarily transpires in those contacts" (41, p. 103). He proposed that the job of the personnel department includes "guidance of the supervisors in the handling of their responsibility for assignment of jobs, evaluation of employee performance and personnel planning" (41, p. 99). Specifically, in regard to employee evaluation Wadsworth stated the following:

> Supervisor's impression of an employee is seldom developed as precisely as should be desired without some prompting. His opinion may, in actuality, be based only upon casual impressions, colored in some degree by his personal reactions to the employee as an individual. . . . That

> our ideas can be wrong is frequently demonstrated when we are called upon to say explicitly where, when, and under what circumstances a given impression has developed. Even the most fair-minded supervisor is not very likely to check himself, or to look for evidence which might revise his opinion of an employee, unless something happens to prompt him to do so. (41, p. 135)

Wadsworth proposed a pattern of inquiry and questioning which begins with an overall evaluation of performance on a three point scale. This rating is followed by a series of suggested probe questions designed to elicit evidence in support of the rating.

Shaeffer, in a general discussion of merit rating plans which included a comparison of methods, stated that "the field review rating procedure is by far the most flexible, comprehensive, and practical rating procedure" (27, p. 698).

Balch, in an article mentioned previously, indicated he had recently heard "Walter Mahler, psychologist and consultant, present a research report on the relative effectiveness of various appraisal methods in which the field review method scored the highest for over-all effectiveness" (1, p. 13).

The Chrysler appraisal system appears to deviate from the field review method as outlined by the above authors in the following respects:

1. Items based on job responsibilities are rated individually
2. Supporting facts are elicited before a rating is obtained

3. The rating is based on a five point graphic scale
4. The overall performance rating is obtained subsequent to item ratings.

Pertinent Research

A search of the literature by the investigator did not reveal any studies that have been published regarding the effect of the coordinator upon the appraisal. Two studies dealing with supervised ratings which appeared to be relevant to the present problem have been published.

In the first study by Taylor and Manson the evaluations were made in the presence of and under direct supervision of a qualified personnel technician. The technician worked with one supervisor at a time and after the preliminary introduction,

> the technician read the first trait description aloud and discussed it briefly as it applied to the specific position. . . . Then the technician asked the supervisor to examine the list of subordinates to be rated and to indicate which was his best worker with respect to that characteristic. When the individual was named the supervisor was asked to justify the choice and indicate where on the scale the rated individual belonged. (36, p. 508)

This study was conducted with five populations of office employees varying in size from 54 to 613 persons. All groups received the same treatment with no control population. The authors evaluated their results in terms of reliability, halo, and distribution skew. Inter-rater reliability coefficients were presented for seven common factors on two of

the groups. The authors stated that these reliability coefficients compared favorably with most found in the literature and were more than adequate for validation purposes. Halo was measured by intercorrelation among traits. The authors said that these coefficients were on the whole considerably lower than those generally found in the graphic scale and so assumed that a step had been made to overcome the problem of halo. Distributions of overall ratings plus means and standard deviations for trait ratings were presented. Excessive skew was not evident as the distributions showed no pile-up in the upper ranges. The authors concluded that supervision of ratings did improve the ratings in the above respects.

This investigator considered that the comparisons made by Taylor and Manson were gross. The lack of a control group and any tests of significance caused this experimenter to question the results.

The second study by Taylor and Hastman (38) was published after the data collection for the Chrysler study had been completed. This study incorporated more in terms of controls and statistical comparisons than the earlier one. The population of 712 office employees was divided among four conditions of administration. These were (1) traditional graphic ratings, (2) Stevens-Wonderlic, (3) Group Stevens-Wonderlic, and (4) Supervised Stevens-Wonderlic. The Stevens-Wonderlic is a practice suggested by them (33) which stated that each trait should occupy a page by itself and each

individual in a group of ratees should be rated on the first trait before considering the next, etc. The evaluations were made on a ten point scale for eight variables common to all four treatments. The relative merits of the four methods were inferred by comparison of what the authors called "secondary criteria of rating system qualities". These were inter-rater reliability, halo, variability, and leniency. As in the previous study the comparisons were primarily on the basis of observation of the results with a lack of tests for significance. The results were negative; no differences were indicated for either format or methods of administration. For all four conditions the findings were:

1. Almost the entire range of ratings was used by the raters
2. The standard deviations indicated a desirable dispersion
3. Rather than peaked, the distributions were somewhat platykurtic
4. Means of the scales were only slightly above the mid-point of the range
5. Only slight negative skewness and no marked asymmetry
6. Correlations among the traits were generally considerably lower than are usually found in graphic rating scales.

The authors considered, "These are rather unusual findings" (38, p. 204). They suggested, by way of explanation,

that these findings may be affected by the following three factors:

1. The ratings were done for research purposes
2. The situation was conducive to the production of desirable ratings
3. The rating scale was well defined in behavioral terms rather than trait names.

Rationale

The purpose of coordination is to increase the efficiency of the appraisal method. This study was directed toward evaluating the influence of the coordinator upon the appraisal. In other words, to what extent, if any, did the presence of a coordinator affect the accuracy or usefulness of the appraisal? The important question then became: How is improvement in a rating method determined? A discussion of the criteria of rating excellence seems appropriate at this point.

As in the case of the program objectives, there are two goals in appraisal:

1. Evaluation or measurement of present performance
2. Improvement of individual performance or planning for personal development.

In order to clarify the problems involved, the two aspects, measurement and development, were considered separately in the following discussion.

Criteria of Evaluation Method

Over thirty years ago Freyd made the statement:

> Ratings are ultimate things, and the comparison of the various systems cannot be found by recourse to an external criterion. In the writer's opinion there are no flawless methods of evaluating rating scales. The criteria which have been advanced may be divided roughly into those which appeal to such factors as ease of administration and scoring, popularity, and so forth, and those which employ statistical reasoning. (12, p. 89)

In regard to the non-statistical criteria he stated, "These are important criteria unless one has access to trained judges with unlimited patience."

Freyd listed the following seven statistical criteria:

1. Comparison of ratings with intelligence test scores
2. Ratings on the same men by the same judge for different months
3. Ratings on same men by different judges
4-5. The form of the distribution -- its normality and its spread
6. Absence of halo
7. Have a person other than the rater sort the ratings and indicate to whom they apply.

The above were not definitive criteria and actually the situation has not changed significantly since that time. Likewise, as was indicated above in 2 and 3, some confusion still exists as to what is reliability and what is validity for a rating method.

Reliability. Two methods of determining the reliability of ratings are commonly used:

1. Comparison of re-ratings by the same rater
2. Agreement between independent raters or inter-rater reliability.

These methods present certain problems. In the first case, re-ratings close together in time frequently produce high reliability coefficients because the rater recalls the previous rating and desires to be consistent. If a long interval of time elapses between ratings a difference in the ratings may reflect a true change on the part of the ratee. The second method assumes that each of the raters is equally familiar with the performance of the ratee and that they assign ratings independently of each other, insuring discreteness not only in rating procedure but also in content. In an industrial situation this is usually not true. Driver stated regarding this method:

> The problem as to whether agreement between a number of independent raters indicates validity or reliability is one that has caused much of this confusion. Some investigators feel that if three or more raters all agree as to a particular quality, this indicates that they are all rating the same trait in the performance they are observing, and hence, that this method is a measure of validity. Others have argued that agreement between judges does not mean they are all drawing the correct conclusions, but that they might agree on some erroneous answer, and that this method is a measure of consistency of rating rather than validity.
>
> This problem is difficult to resolve, but in most investigations such agreement is desired. Whenever individual raters do agree, it is felt that the ratings are of greater value. (7, p. 191)

Whitla and Tirrell (42) in a study concerned with the validity of ratings at various levels of supervision found:

1. The level of raters closest to the ratees was best able to rate them
2. The level of raters closest to the ratees, and only this level, was able to discriminate between sections of the rating instrument.

In the present study it was not possible to obtain ratings for either re-rating or inter-rater reliability. The study was conducted as part of an ongoing program and the management felt that immediate re-ratings would create undue confusion and misunderstanding. Organizational structure was such that the second level of supervision was not close enough to the ratees for rating purposes.

Validity. The problem of validity is the most critical and the most difficult. A reliable rating may not be an actual picture of performance. Driver listed the following as methods employed to obtain validity measures, with varying degrees of success.

Methods of Determining Validity

1. Comparison with some direct measurement of performance, i.e., production records, etc.
2. Comparison with psychological tests purporting to measure the same ability.
3. Comparison with work-samples.
4. Analysis of distribution of results.
5. Analysis to determine the presence or absence of 'halo effect'.
6. Follow-up procedures.
7. Miscellaneous methods. (7, p. 185)

In reference to numbers 4 and 5, he mentioned the dangers involved in making the assumptions necessary for their use. However, he stated that while these methods do not furnish conclusive proof of validity they are considered by many investigators to be fairly good indicators.

Mahler suggested:

> Among the methods of testing the adequacy of the ratings the following are most common:
> 1. Reliability - how consistent are the ratings?
> 2. Validity - how accurate are the ratings?
> 3. Distribution - do the ratings result in adequate spread and does it tend to a normal distribution?
> 4. Halo effect - is there a tendency for the ratings of one or more traits to influence the ratings given to other traits?
> 5. Inter-correlation between traits - do the ratings on different traits tend to be discrete?
> 6. Variation in average rating - is there a wide variation in the average rating of individual raters, the average rating of employees in different occupations? (22, p. 318)

Patterson had the following to say about objective measures, "Many apparently objective criteria are not objective at all. What is usually meant by objective is quantitative. But the most subjective judgments can be expressed in an arbitrary quantitative form" (25, p. 277). He further indicated that military rank is not a valid index of success. Salary is not necessarily related to ability or output. Both output and salary are frequently dictated by union rules.

Taylor and Hastman, on the question of validity, said,

> The most important criterion of the format or administration of a rating scale would thus be its validity. But ratings are themselves frequently the criterion against which predictors are validated. To validate the ratings would require construction of a more ultimate criterion. If such an ultimate criterion could be quickly and economically constructed there would, of course, be no justification for the use of more remote measures of performance such as ratings. (38, p. 184)

Below are listed a few random references and the type of criterion employed to validate rating methods.

Ferguson (9)	- paired comparison ratings
Mahler (24)	- volume scales
Stockford and Bissel (34)	- rating of second supervisor
Chi (5)	- consensus ratings
Taylor, et al. (37)	- salary after bonus proved unfruitful
Sisson (28)	- consensus of fellow-officers.

It appeared to the investigator that a more comprehensive survey would indicate that the most commonly used criteria found in the literature were another rating method, consensus ratings, and ratings of peers or coworkers. Regarding the latter, Springer (31) found a low positive relationship between the ratings given by supervisors and coworkers. There was higher agreement between pairs of coworkers and pairs of supervision.

The evidence seems to indicate that the picture is not clear regarding suitable criteria for the validation of rating methods.

In this study rankings were obtained following the ratings from the same raters for a limited portion of the sample. The question was, whether this is an index of reliability or validity?

Discrimination or Spread. Another approach to this problem can be made in terms of measurement theory and the basic assumptions involved. In other words, an instrument should satisfy certain fundamental requirements to achieve any measurement at all. There are two characteristics of scales which require further attention. The first involves discrimination; the second involves constant and random errors.

A basic assumption in evaluation is that differences do exist between persons in terms of job performance. In order to measure these differences a scale should produce discriminations between the individuals rated. In an industrial situation this is an important problem. Frequently appraisals are obtained which make no discriminations in the group of employees. In this case correlation breaks down and hence reliability and validity cannot be calculated. Consequently, a basic requirement is some variability in ratings. On the other hand, increased dispersion or variability does not insure reliability or validity for in the extreme case this can be due solely to random error.

In the case of appraisal, it is assumed that the rater is not making haphazard responses but that the judgments are based upon observable facts of performance. The question

which arises is whether increased emphasis on a spread of ratings or more discrimination results in an artificial spread or a more accurate measurement of performance. It is a matter of degree rather than an all-or-none proposition.

Various methods of increasing spread are employed, such as rater training, group supervised ratings, forced distribution, etc. Coordination is more than rater training or supervision because it focuses attention upon a specific individual at the time of the rating process rather than dealing with general principles of rating. It is expected that the coordinator obtains increased discrimination which is a more accurate picture of performance. This is achieved through increased clarification of the scale values and explanation of the items plus the continual probing for supporting facts. In other words, the coordinator assists the appraiser in making a more objective evaluation of the employee being rated. Under these circumstances it is assumed that there is less distortion due to forced discrimination than to lack of variability.

Following Guilford (14) the rating a person receives can be thought of as a summation of the "true" value and an error factor. The error term can be further broken down into types of constant errors and residual or random error. He mentions logical error, contrast error, and proximity error with emphasis on the well known types of errors in ratings such as leniency, central tendency, and halo. Guilford explains in detail an analysis of variance procedure which can be used

to identify leniency, halo and contrast. With these error components isolated it is possible to adjust the ratings to eliminate their biasing effects. This analysis of variance procedure can be used only in a situation where two or more raters evaluate the same group of individuals.

Leniency. The tendency of raters to be generous in the evaluation of their subordinates is one that has been frequently noted in the literature by critics of graphic scales. A scale or method which produces ratings with a mean near the mean of the scale and with ratings distributed symmetrically about the mean is considered more useful than one with a high mean and considerable negative skew.

The assumption that normality is the most desirable distribution can be in error. The true distribution of a trait is not known. However, with a large number of cases a spotty or highly skewed distribution indicates that certain portions of the scale are being neglected by raters, or steps in the scale are not equal in value.

Halo. Halo may be defined as a tendency to rate any given employee on the basis of the rater's overall general impression. This is usually noted by the tendency of presumably independent scales to correlate extremely high with each other. The presence of halo is probably the most common criticism of graphic rating scales. The absence of halo is considered by most authorities of rating methods as desirable. The presence of halo tends to lower the validity

of ratings. Closer scrutiny of the literature regarding this topic indicated that the question was not that simple.

This phenomenon called halo effect was first noted by Wells (35) in 1907. Thorndike (39) noticed this tendency for general impression to spread to specific traits and in 1920 gave it the name "halo effect". Several methods have been suggested to reduce halo in ratings such as training raters and better definition of items. The best known technique for graphic scales is that of placing each trait on a separate sheet and rating all stimuli on each trait before proceeding to the next as suggested by Stevens and Wonderlic (33).

Bingham, however, stated that all halo is not invalid, that there is "a halo which cannot and should not be eliminated because it is inherent in the nature of personality, in the perceptual process, and in the very act of judgment" (4, p. 222).

A complete discussion of this topic was not practical at this place but a few comments were made to identify the problem to some extent. Johnson (17) has presented the most thorough coverage of this topic and its relation to the judgment process.

Lynch, in a discussion of the theory of rating scales, indicated that "more than a slight hint of an underlying psychological position" (20, p. 497) is implicit in the assumptions of appraisal.

> The general impression of a person - which is valued by the Gestaltians as reflecting the nature of the whole - is, to the rating-scale practitioner, a phenomenon to be guarded against. Behavior is reduced to isolated parts each of which is evaluated independently, and then combined into a whole.
>
> Here, again, the underlying assumption that the individual is a summation of distinct response systems is in accord with the behavioristic conception. A person may indeed behave as-a-whole, but specific response groups may be isolated and independently evaluated. The whole is exactly equal to the sum of the parts. (20, p. 500)

This would imply that perhaps the view a person takes regarding halo is to some extent a function of his theoretical position. More than likely the answer is somewhere between these two extremes.

The validity or invalidity, as designated by Bingham (4), of halo is no doubt a function of several variables. One that has been generally overlooked is the variation of the stimuli. Johnson (17) pointed out that the actual independence of the traits is of prime importance in order to determine whether the correlation is due to general impression or objective correlation between the traits. In the treatment of this topic writers have dealt with the range from ill-defined personality traits to well-defined aspects of specific behavior without clearly indicating the importance of this factor. Symonds (35) gave five reasons for a large halo which are also indicative of important variables. The trait or habit:

1. Is not easily observed

2. Is not commonly observed or thought about
3. Is not clearly defined
4. Involves reactions with other people rather than "self-contained" behavior
5. Is one with high moral importance in its usual connotation.

Bingham, however, indicated that there probably is an invalid halo "which marks a judgment as vague and undiscriminating, carelessly recorded when the observer's attention has been focussed, not on the trait in relation to its setting, but on the ground alone" (4, p. 223). It is this kind of halo which the experimenter was interested in eliminating from the rating process.

A variety of methods have been used to measure halo. The most common of these is intercorrelations between traits. Factor analytic studies of ratings such as those by Tiffin (40), and Grant (13) have pointed out the presence of a large general factor which is commonly called halo. Johnson and Vidulich (18), using an analysis of variance procedure suggested by Guilford (14), calculated the halo effect as the variance due to interaction between rater and ratee. These authors were able to isolate a significant halo by comparison of conditions designed to maximize and minimize the halo effect. For purposes of this study the experimenter used an index suggested by Bingham (4) to indicate relative halo effect. This was the proportion of identical ratings on all traits.

Criteria of a Development Method

Whether the appraisal method serves to stimulate individual improvement is equally as important as ratings or evaluation. This, of course, depends upon the goals or emphasis of management. The investigator earlier indicated these to be of equal importance in this program.

A thorough evaluation of the effectiveness of the method as a development tool would entail a comprehensive judgment of the entire program. The question would be, does the program produce more effective employees over an extended period of time? An investigation of this sort involves measurement of employee worth in terms of ultimate criteria. This was desirable but impossible to perform because of situational, economic and individual variables which were beyond the control of this experimenter.

Flanagan (10) has criticized conventional approaches to rating and presented the "critical incident" technique as a more satisfactory method of evaluating personnel. He considers the critical incident method primarily as a development tool. In fact, Flanagan and Burns stated:

> The Performance Record is not a yardstick. It is not a rating method. It is a procedure for collecting the significant facts about employee performance.
>
> These facts are gathered in such a way that they will be of maximum usefulness to supervisors and management both in improving the employee's understanding of the requirements of his present job and in developing his potential for more responsible positions. It is not simply a new form but a new approach. (11, p. 102)

The critical incident method, like the system under study, attempts to obtain written evidence of behavior. Likewise, these facts are summarized for an interview with the ratee. The research by Flanagan and Burns (11) did not involve comparing the critical incident with other systems. Their work was concerned with a classification of incidents and the development of the approach rather than a verification of its usefulness.

Without exception authors of appraisal plans designed to improve individual performance agree that a summary of the appraisal is essential and that in some fashion this material must be given back to the ratee. It is only reasonable that an individual must know his weaknesses or development needs in order to do something about them.

The coordinator plays an important role in assisting development through consistent and thorough follow-up subsequent to the appraisal. He also has an important function at the origination of development planning during the appraisal session. The assumption is made that an effective follow through depends to a large degree upon the original analysis of the individual's needs and whether or not plans were made to effectively meet these needs.

The sequence of events during the appraisal and as indicated on the form are as follows:

1. Supporting facts are recorded to justify each rating
2. A performance summary of strengths and weaknesses summarizes the higher and lower ratings

3. The coordinator and appraiser plan tentative development action to meet these specific needs.

If this procedure is followed more realistic and specific development plans should be the result.

An evaluation of this aspect of the appraisal can give immediate criteria of the method as a development technique. In the extreme case, if no summary or development plans are made, none can be suggested to the employee being considered.

Purpose and Criteria of Present Study

It was the objective of this research to determine the effect of the coordinator on the appraisal with respect to the dual aims of evaluation and development. No studies were found which bore directly on this problem. The studies on supervision of ratings were not conclusive. Research aimed at the appraisal as a development instrument has been primarily for the purpose of constructing tools rather than comparing their contribution. The experimenter was also interested in determining the utility of the procedure used for evaluating the development aspects of the appraisal.

Evidence has been discussed which indicates that criteria of a rating method are not easily established nor free from question. In this study no effort was made to compare the methods in terms of outside criteria. The validity and usefulness of the coordinated appraisals were inferred from their relative merits by comparing coordinated appraisals with non-coordinated appraisals in terms of "secondary"

criteria of a rating system" and criteria for the appraisal as a development tool. The latter were established by an examination of the function of the appraisal in the complete process. These were:

1. Discrimination or spread
2. Leniency
3. Hale
4. Ratings versus rankings
5. Quantity and quality of supporting facts
6. Quantity and quality of appraisal summaries
7. Quantity and quality of development plans.

STATEMENT OF PROBLEM

Evaluative Aspects

The evaluative aspects are those concerned with the ratings alone, the scale values of the judgments and their relationships. It is expected that coordination of appraisals will result in improved evaluation of performance. This improvement of the ratings will result from an increase in discrimination and a decrease in the constant errors of leniency and halo. Specifically, it is predicted that the coordinated appraisals will differ from the non-coordinated in those respects stated in the hypotheses listed below.

Discrimination or Spread

Hypothesis 1 - The spread of overall ratings or the use of the scale categories other than the central scale value is greater for the coordinated appraisals.

Hypothesis 2 - The variability of overall ratings for each appraiser is greater for the coordinated group.

Leniency

Hypothesis 3 - The mean of overall ratings for the coordinated group is nearer the central scale value.

Hypothesis 4 - The distribution of overall ratings is more symmetrical, with about the same number of high and low ratings for the coordinated group.

Halo effect

Hypothesis 5 - The variability of item ratings is greater for the coordinated appraisals.

Hypothesis 6 - The coordinated appraisals will contain a smaller proportion of identical ratings.

Coverage

Hypothesis 7 - The coordinated appraisals are more complete and hence contain fewer omitted or "don't know" responses.

Comparability between ratings

Hypothesis 8 - The correlation between separately obtained rank score ratings and overall ratings is greater for the coordinated appraisals.

Development Aspects

The supporting facts or evidence recorded for each item is the basic source of information for determining the performance strengths and needs for each appraisee. A summary of strong points is considered necessary for rounding out the feed-back interview. The summary of needs is the source of material for development planning. This plan should state explicitly how these needs can best be met, the method of training and the responsibility of the employee and/or supervisor for action.

It is expected that coordinated appraisals will contain more adequate supporting facts, more adequate summaries of strengths and needs, and more meaningful development plans than the non-coordinated. The following predictions are

made in regard to the supporting facts and performance summary.

Supporting Facts

Hypothesis 9 - The supporting facts of the coordinated appraisals will:

a. contain a greater amount of information

b. be more descriptive of specific performance

c. contain more criteria of performance

d. contain more examples or illustrations of specific instances.

Performance Summary

Hypothesis 10 - The non-coordinated group of appraisals will contain a greater number of appraisals with sections 1, 2, and 3 completely omitted.

Hypothesis 11 - The coordinated appraisals will contain a larger number of performance strengths (section 1).

Hypothesis 12 - The coordinated appraisals will contain a larger number of development needs (section 2).

Hypothesis 13 - The coordinated appraisals will contain a larger number of methods of handling the development needs.

Hypothesis 14 - The coordinated appraisals will more frequently indicate action to be initiated by the supervisor or both supervisor and employee, rather than employee only.

Hypothesis 15 - On-the-job counselling or coaching is suggested more frequently in the coordinated appraisals as a method of training.

Hypothesis 16 - The development plans of the coordinated appraisals are more specific or concrete than those of the non-coordinated appraisals.

EXPERIMENTAL PROCEDURE

Population

The study was conducted at C.O.N.I.R.A.P. (Chrysler Operated Naval Industrial Reserve Aircraft Plant) formerly known as Chrysler Jet Engine. The plant was operating on a large scale job shop basis, with only limited production line type activity. The employee population represented a wide range of skills and jobs.

The specific group involved in the study could be considered from two points of view:

1. A population of ratees, or
2. A population of raters.

The latter was considered as the unit for the basic analysis.

A sample of 192 persons was appraised or rated by a group of 64 first-line supervisors. First-line supervisors are those who supervise non-management employees, either hourly or salaried.

Each supervisor appraised three persons randomly selected from his work group. The sample of supervisors or appraisers was the starting point for the design layout. The total number of first-line supervisors on the day shift was approximately 70, on the afternoon shift about 10. The sample of 64 supervisors used included all of the total group who satisfied the criteria essential for the study which were

as follows:

1. A supervisor must have three or more persons in his work group (design requirements)
2. The persons appraised must have been under the present supervisor a minimum of three months.

Two of the original group of supervisors left the Corporation before appraisals were begun. Two others were not able to participate because of extremely large groups and extenuating circumstances at the time. These were replaced by supervisors on the afternoon shift.

As the design below indicates, the supervisor sample had to be in multiples of eight in order to simplify calculations. A total of eight conditions or treatments was used and it was desirable to have groups of equal size.

Two of the supervisors were women with work groups consisting of only women employees. An additional ten women employees were present in the other groups giving a total of sixteen.

Design

The basic design involved a division of the entire supervisory sample into two groups:

1. Experimental group or coordinated appraisals
2. Control group or non-coordinated appraisals.

A further breakdown of these two main groups by office vs. plant and by coordinators resulted in the design shown in

Figure I appropriate for a 2 by 2 by 2 analysis of variance factorial design.

Coordinated Group (Experimental)		Non-coordinated Group (Control)	
	Plant		
Coordinator A	Coordinator B	Coordinator A	Coordinator B
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
	Office		
Coordinator A	Coordinator B	Coordinator A	Coordinator B
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-

Figure I. Experimental Design

The above design consisted essentially of eight treatments with eight supervisors in each group. Each of two coordinators coordinated appraisals with 16 supervisors and contacted 16 supervisors in the non-coordinated group. Half of each coordinator's work was with plant supervisors and half with office supervisors.

Sample Selection

The group of 64 supervisors, 32 office and 32 plant, were arranged in pairs according to the function of their work group. This gave two lists of raters in the office and two lists of raters in the plant matched according to the type of work performed. Some of the functional pairs of supervisors were then reversed so that each list contained about the same number of large and small work groups. The supervisors for each location were then assigned in a random fashion to the coordinated or non-coordinated group and to one of two coordinators. See Figure X, in the appendix, which presents the final listing of supervisors by function and group size. The condition and coordinator to which each was assigned is also shown.

The design called for three appraisals from each supervisor. Random numbers were used to select the ratees from an alphabetical listing of personnel in each supervisor's group.

Instrument

The appraisal form presented in appendix Figure XI was prepared for this study. It represented a modification of the management appraisal form. The sequence of procedure and the content required were basically the same as for the management form. The significant changes were in terms of items. The items represented fairly general, yet typical, types of performance rather than managerial responsibilities. No attempt was made to justify the selection of these

particular items. Perhaps it would have been desirable to prepare a scale oriented to more specific job duties but the heterogeneity of the ratee population made this unfeasible. The set of instructions, also shown in appendix Figure XII, served as an outline of appraisal procedure for the introductory sessions and the non-coordinated group of supervisors. The instrument used was common to both the experimental and control groups with differences in administration as the critical variable.

Appraiser Training

At the outset, all supervisors were given an introduction to the program and instructions in appraising through supervisory conferences. These sessions lasted about one hour each, with groups of from 8 to 15 supervisors. A set of instructions and an appraisal form were distributed to each supervisor and the experimenter went through the steps of completing the form and then answered questions which came up.

The majority of appraisals, both coordinated and non-coordinated, were completed in a period of three months although complete collection of the data covered a period of five months.

Coordinators

Two coordinators were used in the study. The experimenter served as a coordinator and was assisted by another

person from the Central Personnel Staff. Neither coordinator had had previous experience as a coordinator. In order to compensate for this fact, both were given opportunities to observe experienced coordinators at work. Both coordinators had attended workshops on the technique of coordination given by Central Personnel Staff for the training of plant coordinators. Following this, the coordinators spent several hours discussing technique and establishing rough ground rules of operation.

Before beginning work with the supervisors in the sample, each coordinator conducted an appraisal while the other observed. In this way it was possible to establish uniformity in approach and begin work with the sample group at a high level of efficiency. During the second week of operation each coordinator was observed at work by the Management Development Program Assistant to determine if his method of coordination was in line with that typically performed.

Administration

The typical pattern was the coordination of three appraisals during a three hour session with a supervisor. The three appraisals were usually completed within this time. The appraisals were conducted away from the supervisor's work area, in a private office.

The coordinator questioned the supervisor about the performance of each ratee in regard to each item. He then

recorded the evidence on the appraisal form and asked the supervisor to rate the item on the five point scale. The supporting facts and the rating were both elicited from the supervisor. The coordinator might have questioned in either case for more detail or evidence. The three ratees were all considered in respect to each item before going on to the next item, etc. The performance summaries were written following the same pattern of inquiry. At the close of the coordination period each supervisor was asked to rank the persons in his group on overall performance.

The non-coordinated supervisors picked up the materials at the office at which time they were given the opportunity to ask individual questions about the procedure. At the time the non-coordinated appraisals were distributed a tentative date was set as to when the supervisor would return them. A week to ten days was the average length of time set. When the supervisor returned his completed appraisals he was asked to rank his entire group and estimate the time spent in making the appraisals.

Very few supervisors completed and returned the appraisals during the initial period of time agreed upon and so it appeared fruitful to record the number of calls made to each person and the amount of time which elapsed between distribution and return of the appraisals for each supervisor.

Data Collection

Each supervisor was assigned a code number. As the appraisals were collected each was marked with the supervisor's number and an employee number. The content of each appraisal was typed in order to expedite coding and to eliminate identification of procedure which might have been possible through handwriting. The typed copy contained only the appraisal number as identification.

Coding

In order to evaluate the content of the written item responses and performance summary content codes or scales were developed by two methods:

1. Apriori, based on assumptions stated as hypotheses
2. Empirically, by an analysis of several appraisals.

The scales were established by the experimenter. The coding was conducted by a person trained for this purpose. The coder was unacquainted with the specific experimental aspects of the study. She was aware that the appraisals were administered under various conditions but did not know what these were nor which appraisals were conducted under a particular condition.

For each scale the experimenter and coder established a common frame of reference by working together on a few appraisals. After definitions and rules for approach were established the coder evaluated the entire group of appraisals.

The experimenter evaluated one-third of the appraisals independently for each scale which served as a reliability check for the coding process.

The item response was the basic unit for the content scales. These were then averaged to obtain a mean scale value for each appraisal. These latter values were the scores used to compare the appraisals in the final analyses.

Statistical Tools

Product-moment correlation, Chi-square, and analysis of variance served as the basic statistical tools for analysis of the data presented.

Chi-square is a non-parametric test which is very useful in comparing distributions where little is known about the form. Hoel stated that "the X^2 distribution is concerned with the values of the e_i but not with the form of the distribution from which they might have been obtained as samples" (15, p. 189). The primary concern with e_i is that no expected cell frequency is less than five. In cases where this does occur other cell frequencies may be combined until the condition is satisfied or corrections made for this discrepancy. In using X^2 it was possible to make the required corrections. Edwards (8) suggested correction for continuity is very easy and should always be made in a 2 by 2 table. This suggestion was followed.

In a discussion of the assumptions underlying the F-test, Lindquist stated, "It is very important, in any application of the simple-randomized design, to consider very carefully the assumptions underlying the F-test of the null hypothesis and the effects on the validity of this test of the failure to satisfy one or more of these assumptions". He further stated, "Generally, if one or more of the conditions is not satisfied, the distribution of ms_a/ms_w will be more variable than the F-distribution. This means that if a 'significant' mean square ratio is obtained in an experiment, it could have resulted from a failure to satisfy any one of these conditions" (19, p. 72).

Bartlett's test of homogeneity of variance was applied in each instance before using analysis of variance to check the assumption that "the variance of the criterion measures is the same for each of these treatment populations". Lindquist stated that this test is only needed when the treatment groups are quite small (probably three or four) or when inspection of the data indicates heterogeneity is marked.

In a few instances the Bartlett's test was significant. A transformation of the data resulted also in a significant Bartlett's test. The F-ratio was significant. According to Lindquist we have occasion to doubt the validity of the F-test when this situation occurs.

The Norton studies on the "Effects of Non-normality and Heterogeneity of Variance" reported by Lindquist (19) served as a reference for the subsequent decision by the experimenter. Norton constructed card populations of 10,000 cases each to empirically study the effects of the above factors. Based on the results of Norton's studies Lindquist made the following statement:

> Accordingly, where marked (but not extreme) heterogeneity is expected, it is desirable to allow for the discrepancy by setting a slightly higher "apparent" level of significance for this test than one would otherwise employ (the "apparent" level being that indicated by the F-table). For example, if one wished the risk of a Type I error to be less than 5%, he might require that the obtained F exceed the 2.5% point in the normal-theory F-distribution. The "apparent" level of significance would then be the 2.5%, but the actual level would be the 5% level. (19, p. 83)

It appeared safe to assume significance in the above mentioned situation, where the F value is not borderline but sufficiently large to allow for any discrepancy which might possibly be due to heterogeneity of variance. In the data reported here the significant F values were considerably larger than double the table values as suggested by Lindquist. The experimenter cites the above as justification for accepting the results in a few such situations as good evidence of significant differences between treatments.

The other alternative was to use a non-parametric test. The use of such a test would prohibit the use of factorial design and hence reduce considerably the amount of information regarding subgroups.

RESULTS AND DISCUSSION

The results of the experiment are presented according to the numerical order of the hypotheses. Hypotheses 1 through 8 are concerned with the evaluative aspects. The development aspects of the appraisal are considered in Hypotheses 9 through 16. Findings dealing with the operational aspects were not outlined by hypotheses. The results are followed by a general discussion.

The statistical design for the experiment was shown in the experimental procedure, Figure I (Page 35). Although this design was a 2 by 2 by 2 analysis of variance factorial resulting in eight treatments, the primary purpose of the study was to determine the effect of the coordinator upon various aspects of the appraisal. A consideration of differences between coordinators and between locations was of secondary interest. The majority of the analyses were conducted by analysis of variance. The results are presented with the emphasis upon the differences between the coordinated and non-coordinated groups. In a few instances where chi square was used for analysis the only comparison was between the coordinated and non-coordinated groups.

The within-groups mean square was used as the error term or the denominator of the F ratio in all cases.

Following a suggestion by Edwards (8, p. 182), no F values were computed if the value of the numerator of the F ratio was smaller than that of the denominator. Therefore, only the F values larger than one are given in the analysis summaries. This procedure facilitates inspection of the results presented in the tables.

Evaluative Aspects

No attempt was made to compare the coordinated appraisals to outside criteria of rating excellence. The tenuousness of such criteria and the reasoning behind this decision were discussed in the introduction. Instead, the relative merits of the system were inferred by comparing the coordinated with non-coordinated appraisals on characteristics of ratings considered desirable or essential.

Discrimination or Spread. It is necessary in any system of evaluation or measurement that distinctions are made between the individuals or stimuli in question. The failure to discriminate between the persons being rated has been a problem with many rating methods. This lack of discrimination is usually indicated by a pile-up of ratings in one category or portion of the scale. Because this concentration of ratings is most frequently found in the middle of the scale it is also called central tendency. Hypotheses 1 and 2 were directed toward this problem.

Hypothesis 1 - The spread of overall ratings or the use of the scale categories other than the central scale value is greater for the coordinated appraisals.

A comparison of the overall rating distributions for the two groups was made by chi square. These data are shown in Table 1. The value of chi square was not significant; consequently, the hypothesis was not confirmed. Inspection of the table indicates there were fewer ratings in the central category for the non-coordinated group, whereas the coordinated group had fewer ratings in categories 3 and 4 combined. This was a rather gross method of comparison but it can be effective in detecting large differences between distributions. It was possible that spread or discrimination measured in this fashion could be a function of rater differences rather than differences in ratee performance.

TABLE 1

COMPARISON OF OVERALL RATING DISTRIBUTIONS FOR THE COORDINATED AND NON-COORDINATED GROUPS

	Category					Total
	1	2	3	4	5	
Coordinated	0	13	51	29	3	96
Non-coordinated	1	7	46	39	3	96
Total	1	20	97	68	6	192
$x^2 = 3.435$						

For 2 df: .05 level $x^2 = 5.991$, .01 level $x^2 = 9.210$

Ideally, spread or discrimination should also be considered within each rater. In other words each appraiser should indicate greater differences between his people for the coordinated group. This was the basis for the next hypothesis.

Hypothesis 2 - The variability of overall ratings for each appraiser is greater for the coordinated group.

Each appraiser rated three persons selected at random from his work group. If discrimination was increased between the ratees of the individual appraisers, the variability of these three ratings should be greater for the coordinated group. The index of variability used was the standard deviation. These data are presented in Table 2. No significant differences were found; therefore, Hypothesis 2 was not confirmed. The coordinated and non-coordinated distributions did not differ significantly in regard to discrimination or central tendency.

Leniency. A scale or method which produces ratings with a mean near the mean of the scale and with ratings distributed symmetrically about the mean is considered more useful than one with a high mean and considerable negative skew. Both of these factors were considered important. A curve could be skewed considerably without affecting the mean significantly due to the influence of extreme values. On the other hand, a distribution could be symmetrical about some point other than the mean. These factors were considered in Hypotheses 3 and 4, respectively.

TABLE 2

ANALYSIS OF VARIANCE OF THE STANDARD DEVIATIONS OF OVERALL RATINGS FOR EACH APPRAISER

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	1,903	1	1903	2.05
Coordinators	75	1	75	
Locations	1,651	1	1651	1.78
Interactions:				
Methods x Coordinators	405	1	405	
Methods x Locations	3,437	1	3437	3.71
Coordinators x Locations	293	1	293	
Methods x Coordinators x Locations	2,104	1	2104	2.27
Within groups	51,932	56	927	
Total	61,800	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

Hypothesis 3 - The mean of overall ratings for the coordinated group is nearer the central scale value.

These data are shown in Table 3. The hypothesis was not confirmed.

TABLE 3

ANALYSIS OF VARIANCE OF THE MEAN OVERALL RATINGS FOR APPRAISERS

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	41	1	41	2.28
Coordinators	51	1	51	2.83
Locations	41	1	41	2.28
Interactions:				
Methods x Coordinators	6	1	6	
Methods x Locations	21	1	21	1.17
Coordinators x Locations	1	1	1	
Methods x Coordinators x Locations	21	1	21	1.17
Within groups	1027	56	18	
Total	1209	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

Hypothesis 4 - The distribution of overall ratings is more symmetrical, with about the same number of high and low ratings, for the coordinated group.

Table 1 contains the data for a comparison of the overall rating distributions in terms of symmetry. The distribution

for the coordinated group was more symmetrical. It had more ratings in the 2 category and fewer in the 4 category than the non-coordinated group. However, both distributions were somewhat skewed and the differences between them were not significant. Hypothesis 4 was not confirmed. The appraisal groups did not differ on the above indexes of leniency.

Halo Effect. Hypotheses 5 and 6 were both directed toward the problem of relative halo. The experimenter felt that these hypotheses were getting at different aspects of the same factor. As indicated in the introduction a variety of measures have been used in the past to indicate the presence or absence of halo. The method used by Johnson and Vidulich (18) was the first to accurately measure and isolate invalid halo. This method required that several raters rate the same stimuli and so was not feasible in this study. The index which was used was based on the proportion of identical ratings as suggested by Bingham (4).

Hypothesis 5 - The variability of item ratings is greater for the coordinated appraisals.

Hypothesis 6 - The coordinated appraisals will contain a smaller proportion of identical ratings.

The data are presented in Tables 4 and 5, respectively. Neither hypothesis was confirmed. The coordinated ratings did not differ from the non-coordinated ratings in the amount of halo present.

TABLE 4

ANALYSIS OF VARIANCE OF THE MEAN OF THE STANDARD DEVIATIONS OF ITEM RATINGS

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	127	1	127	
Coordinators	105	1	105	
Locations	638	1	638	3.69
Interactions:				
Methods x Coordinators	3	1	3	
Methods x Locations	2	1	2	
Coordinators x Locations	10	1	10	
Methods x Coordinators x Locations	232	1	232	1.34
Within groups	9,683	56	173	
Total	10,800	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

TABLE 5

ANALYSIS OF VARIANCE OF THE MEAN PROPORTION OF IDENTICAL RATINGS USED BY AN APPRAISER

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	169	1	169	1.69
Coordinators	10	1	10	
Locations	121	1	121	1.21
Interactions:				
Methods x Coordinators	2	1	2	
Methods x Locations	11	1	11	
Coordinators x Locations	9	1	9	
Methods x Coordinators x Locations	5	1	5	
Within groups	5609	56	100	
Total	5936	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

Coverage. Basic rules of rating state that valid ratings cannot be obtained unless there is an opportunity for the behavior to occur and an opportunity for the rater to observe it. Therefore, rather than force a rater to use every item the directions for appraisal suggested that the

appraiser record "DK" for "don't know" wherever he lacked information on the ratee. The coordinators discouraged the use of this category by questioning directed toward making sure the rater did not know. It was felt that the non-coordinated appraisers would use this category more frequently as an "out" than the coordinated appraisers.

Hypothesis 7 - The coordinated appraisals are more complete and hence contain fewer omitted or "don't know" responses.

Table 6 contains these data. The hypothesis was not confirmed. The non-coordinated appraisals did not contain more "DK" or omitted ratings.

Comparability between Ratings. Subsequent to the appraisal of his people each supervisor was asked to rank the persons in his group according to overall performance on the job. A few supervisors refused to do this ranking because their groups were either too large (30 to 40 persons) or too heterogeneous. The ranks which were obtained were converted to standard scores from a table by Richardson, Bellows, Henry and Company (32, p. 13). This procedure of converting ranks to standard scores is not recommended for groups of less than five persons. Therefore, the ranks for groups of three and four ratees were omitted from this sample. Regardless of these limitations, rank scores on 65 of the coordinated group and 63 of the non-coordinated group were correlated with the mean of the item ratings and the overall

TABLE 6

ANALYSIS OF VARIANCE OF THE MEAN NUMBER OF "DON'T KNOW" OR OMITTED RESPONSES PER APPRAISER. DATA TRANSFORMED TO SQUARE ROOT OF x + .5

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	1	1	1	
Coordinators	20	1	20	
Locations	52	1	52	2.08
Interactions:				
Methods x Coordinators	81	1	81	3.24
Methods x Locations	5	1	5	
Coordinators x Locations	2	1	2	
Methods x Coordinators x Locations	38	1	38	1.52
Within groups	1412	56	25	
Total	1611	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

ratings. The question of whether this was evidence of reliability or validity was not crucial. In either case it was felt that the relationship, and hence the index, should be greater for the coordinated group.

Hypothesis 8 - The correlation between separately obtained rank score ratings and overall ratings is greater for the coordinated appraisals.

These data are shown in Table 7. Differences between the coordinated and non-coordinated groups were not significant. In addition the differences were in the opposite direction to that predicted. Hypothesis 8 was not confirmed. The correlations were all large enough to reject the hypothesis that the population correlation was zero. They were not large enough to be considered evidence for satisfactory reliability, but compare favorably with correlations of this type frequently presented in the literature.

TABLE 7

CORRELATION BETWEEN RANK SCORE RATINGS AND MEAN RATINGS AND BETWEEN RANK SCORE RATINGS AND OVERALL RATINGS

	Mean ratings			Overall ratings		
	n	r	z'	n	r	z'
Coordinated	65	.51	.563	65	.47	.510
Non-coordinated	63	.56	.633	63	.65	.775
		z = .388			z = 1.478	

Differences also in opposite direction to that predicted.

$z_{.05} = 1.96$ $z_{.01} = 2.58$

For 60 df: .05 level r = .250, .01 level r = .325

Item Correlations. A step beyond those factors considered in the above hypotheses was taken to explore further the possibility of differences between the coordinated and non-coordinated groups. This consisted of calculating item-overall and item-potential correlations for the coordinated and non-coordinated groups individually and combined. These correlation coefficients are shown in Table 8. Inspection of the table did not indicate any large differences between the coefficients of the coordinated and non-coordinated groups for items with overall judgments. In addition, neither group was consistently above the other in terms of the size of coefficients. The same comments apply to the item-potential correlations except for one fact. The non-coordinated groups contains a significantly (at .05 level) greater number of not significant coefficients than the coordinated group. The data for this comparison are shown in Table 9. This was some evidence that the non-coordinated group of appraisers did not see the same relationship between the potential rating and these particular items as did the coordinated appraisers. The concept of potential proved difficult for many supervisors to understand. Explanation by the coordinators no doubt helped to clarify the potential rating. It does not follow, however, that increased understanding should result in a stronger relationship of the potential ratings to these items alone.

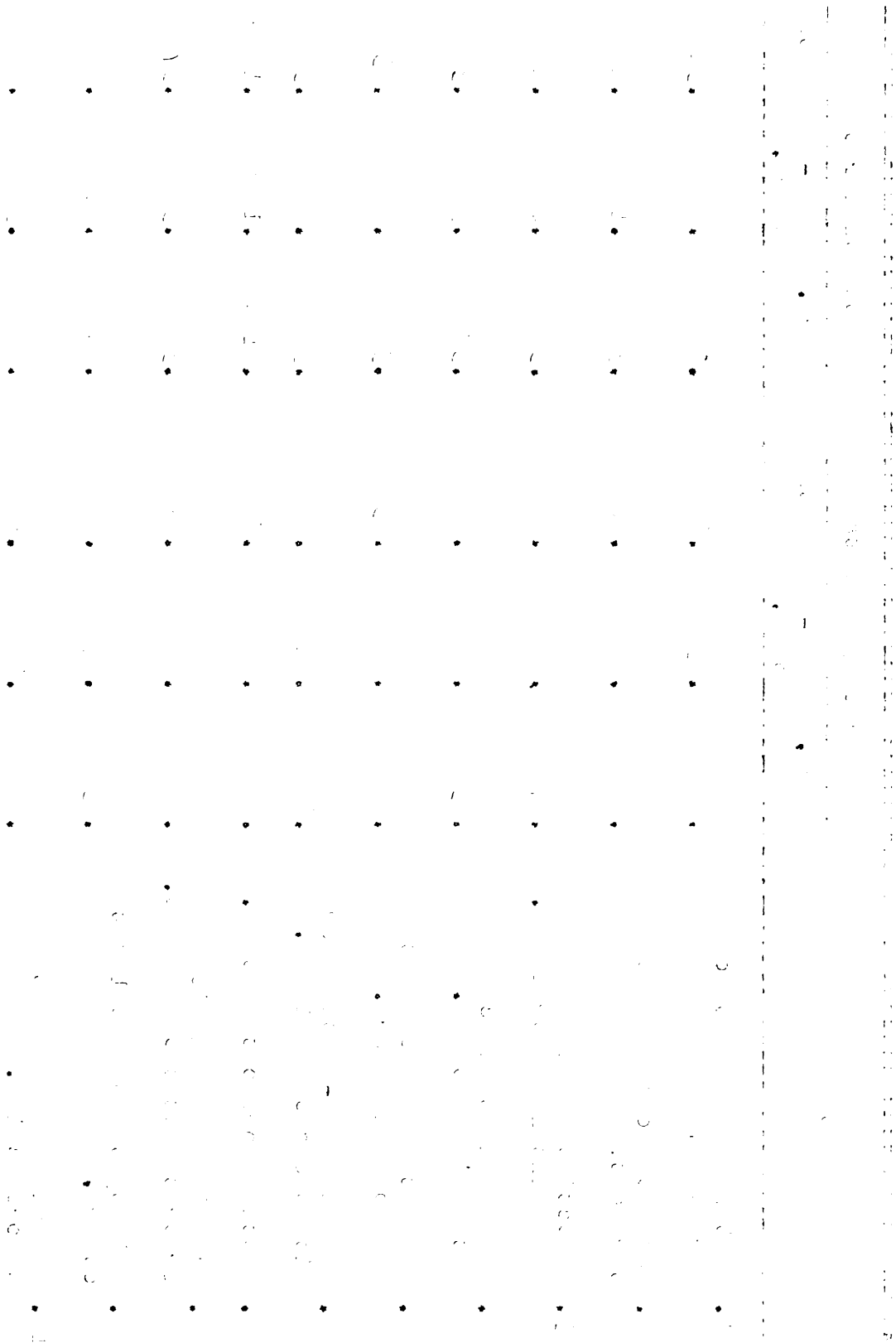


TABLE 8

CORRELATIONS BETWEEN ITEMS, OVERALL JUDGMENTS, AND POTENTIAL RATINGS

Item	Overall judgments			Potential ratings		
	Coord.	Non-coord.	Both	Coord.	Non-coord.	Both
1. What is the quality of work done?	.64	.63	.64	.35	.29	.32
2. How much work does he turn out?	.58	.40	.50	.34	.14 ns	.24
3. Effectiveness in working under little supervision.	.67	.65	.66	.34	.27	.29
4. Dependability in carrying out assigned tasks.	.63	.55	.59	.37	.37	.35
5. Suggesting improvements in ways of doing work.	.42	.50	.43	.39	.27	.33
6. Handling non-routine work situations effectively.	.69	.72	.69	.37	.40	.39
7. Getting work out on time.	.50	.62	.57	.21 *	.15 ns	.18 *
8. Assuming added duties in order to get the job done.	.61	.54	.57	.39	.34	.37
9. Ability to work well with others.	.43	.54	.49	.20 *	.31	.25
10. Thinking a problem through before acting.	.60	.57	.58	.40	.39	.39

11. Observing rules and regulations.	.29	.39	.36	.18 *	.06 ns	.12 ns
12. Efforts to learn about other phases of the work than his own.	.50	.53	.50	.40	.46	.43
13. Following job instructions.	.52	.58	.55	.27	.13 ns	.18 *
14. Speed in learning new duties.	.49	.39	.45	.56	.14 ns	.35
15. Keeping superiors informed on matters of mutual concern.	.48	.52	.51	.22 *	.20 ns	.27
16. Accepting without complaint, adverse work conditions and difficult job assignments.	.30	.56	.45	.18 ns	.33	.25
17. Ability to express himself orally.	.45	.15 ns	.28	.28	.04 ns	.09 ns
18. Ability to express himself in writing.	.39	.31	.35	.39 *	.01 ns	.14 ns
Potential	.50	.45	.47			

Correlations for coordinated and non-coordinated groups are based on n's ranging from 43 to 96.

ns - Not significantly different from zero
* - Significant at .05 level; all others significant at .01 level

TABLE 9

COMPARISON BETWEEN COORDINATED AND NON-COORDINATED GROUPS ON THE NUMBER OF NOT SIGNIFICANT CORRELATIONS OF ITEM RATINGS WITH POTENTIAL RATINGS

	Not significant	Significant	Total
Coordinated	1	17	18
Non-coordinated	8	10	18
Total	9	27	36
$x^2 = 5.332$*			

For 1 df: .05 level $x^2 = 3.841$, .01 level $x^2 = 6.635$
*Significant beyond .05 level

No evidence was found in this study to indicate that the method of coordination of appraisals had any effect on the ratings in those respects stated in Hypotheses 1 through 8. At this point it might appear that these hypotheses were not originated on a sound basis. In order to partially justify the position stated in the hypotheses two items of information are presented and discussed.

Taylor and Manson (36), in a study mentioned earlier, concluded that ratings individually supervised by a personnel technician were more reliable, contained less halo, and showed less skew than non-supervised ratings. In this particular study there was no control group; the comparisons were gross, with no tests of significance. Perhaps the authors reached conclusions not adequately supported by their data.

Regarding the second item, Table 10 shows the overall rating distributions of coordinated appraisals from several Chrysler plants not included in this study. Two things were indicated in these distributions: 1) a definite central tendency, and 2) a leniency tendency (more 4 ratings than 2 ratings). These data were some of the first available at the outset of the appraisal program. Increased emphasis on discrimination by coordinators later improved the distributions considerably. It was expected, therefore, that non-coordinated appraisals would show even more central tendency and/or leniency than that noted in Table 10.

Table 11 shows the distributions for both the coordinated and non-coordinated groups with the percent of responses in each category. A visual comparison of the percentages in each category indicated less central tendency and leniency in both the distributions in Table 11 than those in Table 10. This finding had not been anticipated in the non-coordinated group (Hypotheses 1 and 4).

In Table 12 a comparison was made between the Dodge distribution at the bottom of Table 10 and the coordinated group from this study. A significant chi square was found. Since the groups of this study did not differ significantly, a general improvement over the Dodge distribution was indicated. If this general improvement had not been found in the non-coordinated group, a significant chi square would have been found and Hypotheses 1 and 4 would have been confirmed.

TABLE 10

DISTRIBUTIONS OF OVERALL RATINGS FOR
FIRST LEVEL SUPERVISORS PREVIOUS TO STUDY

Plant	Category					
	1	2	3	4	5	Total
Dodge	2	15	134	41	0	192
Mopar	0	15	58	7	0	80
L.A.	0	31	65	11	0	107
Canada	0	5	123	21	1	150
Total	2	66	380	80	1	529
Percent	.4	12.5	71.8	15.1	.2	100
ABD Mack	0	3	408	77	1	489
Percent	0	.6	83.4	15.7	.3	100
Dodge	0	5	184	65	0	254
Percent	0	2.0	72.4	25.6	0	100

TABLE 11

DISTRIBUTIONS OF OVERALL RATINGS FOR COORDINATED
AND NON-COORDINATED GROUPS FROM PRESENT STUDY

	Category					
	1	2	3	4	5	Total
Coordinated	0	13	51	29	3	96
Percent	0	13.5	53.1	30.3	3.1	100
Non-coordinated	1	7	46	39	3	96
Percent	1.0	7.3	47.9	40.7	3.1	100

TABLE 12

COMPARISON OF OVERALL DISTRIBUTION OF DODGE WITH DISTRIBUTION OF COORDINATED GROUP FROM PRESENT STUDY

	Category			Total
	2	3	4	
Dodge	5	184	65	254
Coordinated	13	51	32	96
Total	18	235	97	350
$x^2 = 22.633^{**}$				

For 2 df: .05 level $x^2 = 5.991$, .01 level $x^2 = 9.210$
**Significant beyond .01 level

In the study by Taylor and Hastman (38) no differences were found among four groups of ratings in regard to alternate rater reliability, dispersion, halo, and leniency. The four groups involved two types of format and two types of administration. One type of administration was individually supervised ratings, which was relevant to this study. These authors attribute their lack of positive findings to the acceptance of the rating system and a situation conducive to the production of desirable ratings. As in this study, the ratings of the control group were more satisfactory on the characteristics measured than was expected.

Development Aspects

A second, but equally important, objective of the appraisal program was the individual development or improvement of job performance for each appraisee. The appraisal instrument served as a record of present performance and of the plans for proposed development action. The results and discussion presented in this section were directed toward the evaluation of the appraisal in this function, and the extent to which the coordinator affected its usefulness as such an instrument. Specifically, the experimenter was interested in the quantity and quality of the:

1. Supporting facts
2. Appraisal summaries
3. Development plans.

The evaluation or classification of written material for a particular content or the detection of degrees of emphasis intended by the respondent to a question is called content analysis or coding. Like rating it is a somewhat subjective technique and involves the scaling or categorization of the written material on the basis of a selected characteristic or factor. It is the same process as rating. The coder or analyst reads the material and evaluates or classifies it according to a specific factor previously established. This method of analysis served as a basic device for the evaluation of the information recorded on the appraisal form. The investigator used "A Manual for Coders" (16) as the primary source of information on the utilization of this technique.

Because of its subjective nature it is recommended practice that the coding be done by a person other than the experimenter. An advanced college student was selected and reimbursed as coder for this investigation. She was given a copy of the manual to study. Time was spent with the experimenter in a discussion of the general principles involved. The coder was not aware of the experimental design or purpose of the experiment. The appraisal material was typed on sheets identified by code numbers known only to the experimenter. The scales were set up by the investigator. The experimenter then went over the scale with the coder to explain and illustrate its particular application. Following this initial practice both the coder and experimenter evaluated independently one-third of the appraisals. This one-third consisted of one appraisal for each supervisor and hence covered all treatments of the design. A correlation was computed between the values assigned the appraisals by the coder and the experimenter. This served as an index of reliability. If satisfactory agreement was evident the coder then evaluated the remaining two-thirds of the appraisals.

In the case of more objective evaluation, such as the counting of words, counting of particular words, and recording of the presence or absence of data, reliability checks were not conducted. In these instances some of the work was done entirely by the coder and some by the experimenter.

The supporting facts were evaluated on the basis of content per item. These item values were averaged to obtain an appraisal score. For the performance summaries and development plans an appraisal score was obtained directly because there was only one occurrence per appraisal.

Supporting Facts. The supporting facts or evidence for the ratee's behavior on a specific item was considered the key to subsequent planning because it was the initial record of details of performance. Hypothesis 9 covered four aspects of the supporting facts.

Hypothesis 9 - The supporting facts of the coordinated appraisal will:

a) contain a greater amount of information
b) be more descriptive of specific performance
c) contain more criteria of performance
d) contain more examples or illustrations of specific instances.

Figure II contains the rules or definitions used to evaluate two aspects of Hypothesis 9a, the amount of information. In both instances these guides were developed by considering a cross section of the raw data. Both of these counts were directed toward the measurement of the amount of material present on the appraisal form.

Table 13 contains the analysis summary for the word count. A very significant F ratio was found between methods. The coordinated appraisals contained more words than the

Word Count

All words counted except:

- Personal pronouns
- Articles
- Conjunctions
- Two-letter words

Thought Count

Major thought -

- Direct or fairly direct answer to the question
- Physically present, not only implied
- Usually contains adjective or adverb stating how person does it
- Only one

Minor thought -

- Examples or illustrations
- Explanatory or qualifying statements - under what conditions, amount of time, etc.
- Conditional statements
- Supports major thought
- None or several

Reliability: Major count, r = .77

Minor count, r = .96

Figure II. Prescribed Guide for Word and Thought Counts

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

TABLE 13

ANALYSIS OF VARIANCE OF THE
MEAN NUMBER OF WORDS PER APPRAISAL

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	85,923	1	85,923	33.97**
Coordinators	7,548	1	7,548	2.98
Locations	10,177	1	10,177	4.02*
Interactions:				
Methods x Coordinators	17,789	1	17,789	7.03*
Methods x Locations	1,234	1	1,234	
Coordinators x Locations	1,269	1	1,269	
Methods x Coordinators x Locations	747	1	747	
Within groups	141,636	56	2,529	
Total	266,323	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

*Significant beyond .05 level
**Significant beyond .01 level

Bartlett's test: Significant at .05 level

non-coordinated. Heterogeneity of variance between treatments indicated by the significant Bartlett's test may have contributed to the significant F ratio. However, the F was large enough to assume that this was not the sole cause for significance. This was the first occurrence where a transformation of data did not reduce the heterogeneity. Norton's studies are referred to in support for the acceptance of such situations.

Table 14 shows the data for the major thought count. No significant differences were found. As the definition indicates, the major thought was any answer to the item. These were identified and counted primarily to allow more accurate isolation and counting of the minor thoughts. The latter were considered a better index of the amount of information and, as defined, the kind of information considered good evidence of performance. Because the major thoughts were such elementary responses it was understandable that no differences were indicated.

The data for the minor thought count are shown in Table 15. A significantly greater number of minor thoughts were found in the coordinated appraisals. In addition, the appraisals done by coordinator B contained more minor thoughts than those done by coordinator A.

As defined by word and minor thought counts, the coordinated appraisals contained a greater amount of information. Hypothesis 9a was supported.

TABLE 14

ANALYSIS OF VARIANCE OF THE
MEAN NUMBER OF MAJOR THOUGHTS PER APPRAISAL

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	877	1	877	1.51
Coordinators	206	1	206	
Locations	819	1	819	1.41
Interactions:				
Methods x Coordinators	1,114	1	1114	1.92
Methods x Locations	21	1	21	
Coordinators x Locations	147	1	147	
Methods x Coordinators x Locations	1,066	1	1066	1.84
Within groups	32,409	56	579	
Total	36,659	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Significant beyond .01 level

TABLE 15

ANALYSIS OF VARIANCE OF THE
MEAN NUMBER OF MINOR THOUGHTS PER APPRAISAL

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	504,278	1	504,278	159.38**
Coordinators	37,977	1	37,977	12.00**
Locations	8,441	1	8,441	2.67
Interactions:				
Methods x Coordinators	50,120	1	50,120	15.84**
Methods x Locations	2,244	1	2,244	
Coordinators x Locations	29	1	29	
Methods x Coordinators x Locations	251	1	251	
Within groups	177,175	56	3,164	
Total	780,515	63		

For df 1 and 55: $F_{.05} = 4.12$, $F_{.01} = 7.12$

**Significant beyond .01 level

Bartlett's test: Not significant

Figure III contains a description of Scale I and an index of its reliability. In Table 16 the analysis summary is shown. The F ratio between methods was large enough to be considered very significant even though Bartlett's test was also significant. The item responses for the coordinated

Scale I

How complete or explicit is the answer?

"What has he done (any evidence or explanation)?"

Categories:

1. Not stated
2. Vague - incomplete
3. Understandable - somewhat complete
4. Clear - complete

Reliability: r = .91

Figure III. Scale I: Definition and Categories

groups were scaled higher in regard to completeness or explicitness. Hypothesis 9b was confirmed.

Figure IV contains illustrations of criteria or dimensions. These illustrations indicate what was considered in this respect. The criteria were selected through examination of several appraisals.

A summary of the analysis is shown in Table 17. Significant F ratios were found for the three main effects and one interaction. There were more criteria present in the coordinated than non-coordinated appraisals, more in the appraisals of coordinator B than A. The interaction meant that one of the coordinators, B, made more use of coordination than the other. Actually neither coordinator had much effect on the non-coordinated appraisals in respect

TABLE 16

ANALYSIS OF VARIANCE OF SCALE I MEAN VALUES.
HOW COMPLETE OR EXPLICIT IS THE ANSWER?

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	506	1	506	25.95**
Coordinators	14	1	14	
Locations	14	1	14	
Interactions:				
Methods x Coordinators	39	1	39	2.00
Methods x Locations	39	1	39	2.00
Coordinators x Locations	.3	1	.3	
Methods x Coordinators x Locations	.7	1	.7	
Within groups	1091	56	19.5	
Total	1704	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

**Significant beyond .01 level

Bartlett's test: Significant beyond .01 level

to this scale. Therefore, a significant F for this interaction only substantiates the difference between coordinators noted in the main effect. Hypothesis 9c was confirmed.

Figure V shows Scale II. The category definitions are self-explanatory and indicate rather clearly what was

Item	Criteria or Dimensions
1 - What is the quality of work done?	Errors - scrap - rework Correction - clear - concise Tolerance - standards
8 - Assuming added duties in order to get the job done.	Emergency Vacation time Out of classification Repairs equipment Trains others
9 - Ability to work well with others.	Co-workers Supervisor Other departments Vendors
11 - Observing rules and regulations.	Absence - tardiness Safety - relief time Accidents - care of tools

Reliability: r = .92

Figure IV. Examples of Criteria or Dimensions of Performance

Scale II

Kind of information or evidence - type of material.

Categories:

1. Only adjectives or adverbs - restatement of item - verbal expression of scale category - personality characteristics or traits
2. Restatement plus qualification - enlargement or explanation in general terms
3. Example by class of work - reference to total skill area - performance on the job
4. Example or illustration by reference to a specific job or occasion. A for-instance.

Reliability: r = .94

Figure V. Scale II: Definition and Categories

TABLE 17

ANALYSIS OF VARIANCE OF THE
MEAN NUMBER OF DIMENSIONS OR CRITERIA PER APPRAISAL

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	120,235	1	120,235	169.35**
Coordinators	8,977	1	8,977	12.64**
Locations	5,814	1	5,814	8.19**
Interactions:				
Methods x Coordinators	11,583	1	11,583	16.31**
Methods x Locations	473	1	473	
Coordinators x Locations	127	1	127	
Methods x Coordinators x Locations	61	1	61	
Within groups	38,766	56	710	
Total	187,036	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

**Significant beyond .01 level

Bartlett's test: Not significant

considered in each. The analysis summary is shown in Table 18. The F ratio between methods was large enough to be considered significant even with a significant Bartlett's test. The coordinated appraisals were found to contain more examples by

TABLE 18

ANALYSIS OF VARIANCE OF SCALE II MEAN VALUES.
KIND OF INFORMATION OR EVIDENCE

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	473	1	473	26.28**
Coordinators	6	1	6	
Locations	39	1	39	2.17
Interactions:				
Methods x Coordinators	18	1	18	1.00
Methods x Locations	100	1	100	5.56*
Coordinators x Locations	1	1	1	
Methods x Coordinators x Locations	1	1	1	
Within groups	1006	56	18	
Total	1644	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

*Significant beyond .05 level
**Significant beyond .01 level

Bartlett's test: Significant beyond .01 level

class of work and examples of specific jobs or occasions than the non-coordinated appraisals. The F ratio for the interaction methods by locations may have been a function of heterogeneity of variance as it was only slightly beyond the .05 level of significance. Hypothesis 9d was confirmed.

The statements in Hypothesis 9a through 9d and the indexes used to check these were established by the experimenter. They were developed from the stated objectives of the appraisal program and the rather global criteria for good supporting facts suggested by persons experienced in appraisal and responsible for the operation of the program and training of new coordinators. The statements or hypotheses may not have represented completely independent factors. No index of their independence was established. They were directed toward what seemed to be different aspects of the quantity and quality of supporting facts. It was not considered as important that each scale measure a specific factor as it was that quantity and quality be investigated thoroughly.

An attempt was made to confirm the fact that these were important factors. This was done by an expert evaluation of the appraisals.

Figure VI shows the scale and directions used for an expert evaluation of the supporting facts. Two experienced coordinators, not involved in the experiment, were asked to sort the entire sample of appraisals using these instructions.

Expert Sort of Item Content

Directions:

What is the amount of information given about a person's performance?

How much does it tell you about what and how the person does?

Consider both:
Quality or specificity
Quantity or coverage

Sort into five stacks then clip the category card on the front of each group and return.

Consider only the supporting facts in Items 1 - 18.

Categories:

1. Very poor - very unsatisfactory - least
2. Poor - less than expected
3. Fair - average - good as can be expected
4. Good - more than expected
5. Very good - very satisfactory - most

Reliability: r = .72, n = 192

Figure VI. Directions and Categories for Expert Evaluation of Item Content

They did not evaluate each item but the complete group of Items 1 through 18. The correlation between these two sorts was taken as an index of reliability and is shown in Figure VI.

An analysis of variance was conducted using the mean score for the two experts. This summary is shown in Table 19. A significant Bartlett's test was found indicating

TABLE 19

ANALYSIS OF VARIANCE OF MEAN SCORES FROM EXPERT EVALUATION, ON AMOUNT AND SPECIFICITY OF INFORMATION

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	52,041	1	52,041	29.47**
Coordinators	18,666	1	18,666	10.57**
Locations	7,119	1	7,119	3.11
Interactions:				
Methods x Coordinators	31,907	1	31,907	18.07**
Methods x Locations	1,000	1	1,000	
Coordinators x Locations	900	1	900	
Methods x Coordinators x Locations	1,515	1	1,515	
Within groups	98,875	56	1,766	
Total	212,023	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

**Significant beyond .01 level

Bartlett's test: Significant beyond .01 level

heterogeneity of variance. The F ratio between methods was large enough to assume significance in light of the Bartlett's test. The coordinated appraisals were rated as highest on both quantity and quality of information given in the supporting facts. The F ratio found between coordinators was not much beyond the .01 level of significance. It showed that the appraisals done by coordinator B were evaluated higher than those by coordinator A. The F ratio for methods by coordinators confirms this difference between coordinators. This finding was considered as a confirmation of the counts and Scales I and II and consequently support for all sections of Hypothesis 9.

An effort was made to determine whether the increased quantity and quality of supporting facts for the coordinated appraisals actually represented better evidence for the ratings given by the appraisers. Four persons were asked to read the supporting facts and assign an overall rating of performance to the ratee based on this information. Correlations were computed between the actual and estimated ratings. The entire group of appraisals was done by the coder and a naive person. Two experienced coordinators each did one-third of the appraisals. The coefficients were computed for each one-third of the appraisals and broken down by coordinated versus non-coordinated. The coefficients and their z' score equivalents are given in Table 20. The difference between these two groups of coefficients was

TABLE 20

CORRELATION BETWEEN OVERALL RATINGS OF THE APPRAISERS AND RATINGS ESTIMATED FROM THE ITEM CONTENT

Person	Coordinated			Non-coordinated		
	n	r	z'	n	r	z'
A	32	.73	.929	32	.63	.741
A	32	.83	1.188	32	.73	.929
A	32	.74	.950	32	.87	1.333
B	32	.81	1.127	32	.56	.633
C	32	.86	1.293	32	.35	.365
D	32	.68	.829	32	.58	.662
D	32	.76	.996	32	.66	.793
D	32	.63	.741	32	.52	.576
		m = 1.007			m = .754	

$$z = \frac{.253}{.111} = 2.279*$$

$z_{.05} = 1.96$, $z_{.01} = 2.58$

For 30 df: .05 level r = .349, .01 level r = .449

significant beyond the .05 level. This indicates that some additional information was found in the content of the coordinated appraisals which would allow a person other than the appraiser to duplicate his rating.

Additional exploratory analyses were conducted on the item content not directly related to the stated hypotheses. Figure VII shows a scale of general tone or affect. The summary data are given in Table 21. No significant F ratios were found. It was suggested that this result may have been distorted by the use of mean values and scores per appraisal. In order to check on this an analysis of variance was conducted of the number of 1 and 2 values and of the number of

Scale III

General tone or affect (feeling).

Categories:

1. Very negative - entire negative
2. More negative than positive
3. Neutral - fifty-fifty
4. More positive than negative
5. Very positive - entirely positive

Reliability: r = .96

Figure VII. Scale III: Definition and Categories

TABLE 21

ANALYSIS OF VARIANCE OF SCALE III MEAN VALUES. TONE OR AFFECT

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	30	1	30	1.35
Coordinators	7	1	7	
Locations	12	1	12	
Interactions:				
Methods x Coordinators	.6	1	.6	
Methods x Locations	42.0	1	42.0	1.89
Coordinators x Locations	.6	1	.6	
Methods x Coordinators x Locations	28.8	1	28.8	1.30
Within groups	1243	56	22.2	
Total	1364	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

4 and 5 values for Scale III per appraisal. These data are shown in Tables 22 and 23, respectively. No significant F ratios were found. In regard to negative and positive responses to the items no differences were found between groups.

TABLE 22

ANALYSIS OF VARIANCE OF THE
MEAN NUMBER OF 1 AND 2 VALUES FOR SCALE III

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	1,038	1	1,038	2.50
Coordinators	1	1	1	
Locations	86	1	86	
Interactions:				
Methods x Coordinators	16	1	16	
Methods x Locations	1,296	1	1,296	3.12
Coordinators x Locations	11	1	11	
Methods x Coordinators x Locations	210	1	210	
Within groups	23,221	56	415	
Total	25,879	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

TABLE 23

ANALYSIS OF VARIANCE OF THE
MEAN NUMBER OF 4 AND 5 VALUES FOR SCALE III

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	2	1	2	
Coordinators	85	1	85	
Locations	663	1	663	
Interactions:				
Methods x Coordinators	110	1	110	
Methods x Locations	256	1	256	
Coordinators x Locations	28	1	28	
Methods x Coordinators x Locations	553	1	553	
Within groups	41,456	56	740	
Total	43,153	63		

For df 1 and 55: $F_{.05} = 4.12$, $F_{.01} = 7.12$

Bartlett's test: Not significant

Figure VIII shows a list of typical words used to modify statements about performance. Examination of the appraisals gave rise to the question as to whether the coordinated appraisals differed from the non-coordinated in the use of such words. The analysis summary for the so-called dogmatic and qualifying words is shown in Tables 24 and 25, respectively. No significant F ratios were found in the first case, indicating no difference in the use of the number of dogmatic words. A significant F ratio between methods was found in regard to qualifying words. Words of this sort were used more frequently in the coordinated appraisals than in the non-coordinated appraisals. This indicated that the appraiser was perhaps more cautious in his statements to a coordinator than when he recorded them himself.

Classes of Modifying Words

Dogmatic	Qualifying
Never	Occasionally
None	Sometimes
Always	Usually
Constantly	Frequently

Figure VIII. Examples of So-called Dogmatic and Qualifying Words

A significant Bartlett's test, which indicated heterogeneity of variance between treatments lowered the precision of the F test. In those cases where this was true a much

TABLE 24

ANALYSIS OF VARIANCE OF THE MEAN NUMBER OF EXTREME OR DOGMATIC WORDS. DATA TRANSFORMED TO SQUARE ROOT OF x + .5

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	5,166	1	5,166	2.08
Coordinators	1,754	1	1,754	
Locations	58	1	58	
Interactions:				
Methods x Coordinators	5,184	1	5,184	2.09
Methods x Locations	1,360	1	1,360	
Coordinators x Locations	3,235	1	3,235	1.30
Methods x Coordinators x Locations	1,704	1	1,704	
Within groups	139,041	56	2,483	
Total	157,502	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

Bartlett's test: Not significant

TABLE 25

ANALYSIS OF VARIANCE OF THE MEAN NUMBER OF QUALIFYING WORDS. DATA TRANSFORMED TO SQUARE ROOT OF x + .5

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	37,153	1	37,153	20.47**
Coordinators	218	1	218	
Locations	2,998	1	2,998	1.65
Interactions:				
Methods x Coordinators	5,184	1	5,184	2.86
Methods x Locations	225	1	225	
Coordinators x Locations	132	1	132	
Methods x Coordinators x Locations	161	1	161	
Within groups	101,645	56	1,815	
Total	147,716	63		

For df 1 and 55: $F_{.05} = 4.12$, $F_{.01} = 7.12$

**Significant beyond .01 level

Bartlett's test: Not significant

greater F ratio was needed to assume the existence of significant differences. This was no problem with these data and particularly in regard to the main effect of methods which was of primary interest. On the other hand the Bartlett's test did give useful information regarding the differences between the coordinated and non-coordinated groups of appraisals. In every case where heterogeneity was indicated this was a function of the reduction in variance found in the coordinated groups. Coordination reduced the variation between appraisers for the various treatment groups on these particular scores. This means that for the coordinated groups the appraisals were more uniform or alike than the non-coordinated. This was true for the following factors:

1. Number of words
2. Number of major thoughts
3. Scale I - completeness or explicitness
4. Scale II - kind of evidence
5. Expert evaluation.

Coordination then assures more consistency between appraisers on these aspects of quantity and quality of the supporting facts.

Performance Summary. The performance summary, Part II, of the appraisal contained three sections: (1) strengths and abilities, (2) development needs, and (3) methods of development. The thorough and detailed use of this part of the appraisal form was considered essential to any subsequent

action. Preliminary observation of the completed forms showed that some appraisers had failed to fill in this part of the appraisal. For this part of the form satisfactory completion appeared to be a contribution of the coordinator.

Hypothesis 10 - The non-coordinated group of appraisals will contain a greater number of appraisals with sections 1, 2, and 3 completely omitted.

Table 26 shows a comparison between the two groups. The three chi square values are all significant beyond the .01 level. Only 14 non-coordinated appraisals lacked a summary of strengths and abilities while 25 did not contain any development needs. This difference may have indicated a reluctance on the part of the supervisor to record what he considered unfavorable or a lack of understanding as to what he should record. He may have considered the employee entirely satisfactory and not in need of improvement.

Thirty-two non-coordinated appraisals lacked development plans. The written instructions for appraisal as well as the introductory oral instructions emphasized the completion of Part II of the appraisal form. In both cases definite suggestions were made regarding the type of material expected in the section on methods of development. It is unlikely that strengths, needs, or plans would be communicated to the ratee in a feedback interview unless these appeared on the appraisal form. Likewise, follow-up with the supervisor on the course of development action would be impossible if no development plans were suggested.

TABLE 26

COMPARISON OF COORDINATED AND NON-COORDINATED GROUPS FOR RESPONSES TO SECTIONS 1, 2, AND 3 OF PART II

	With response	Without response	Total
Section 1			
Coordinated	94	2	96
Uncoordinated	82	14	96
Total	176	16	192
$x^2 = 8.035^{**}$			
Section 2			
Coordinated	89	7	96
Uncoordinated	71	25	96
Total	160	32	192
$x^2 = 10.837^{**}$			
Section 3			
Coordinated	94	2	96
Uncoordinated	64	32	96
Total	158	34	192
$x^2 = 30.057^{**}$			

For 1 df: .05 level $x^2 = 3.841$; .01 level $x^2 = 6.635$

**Significant beyond .01 level

The following analyses concerning the number of strengths, needs and methods of development exclude the appraisals where no material was recorded. This was done to eliminate the bias caused by the appraisals where the sections were omitted.

Hypothesis 11 - The coordinated appraisals will contain a larger number of performance strengths (Section 1).

Table 27 contains these data. Hypothesis 11 was confirmed. The F ratios for locations and methods by locations were barely significant at the .05 level. Because the Bartlett's test was also significant at the .05 level these were not considered large enough for significance.

Hypothesis 12 - The coordinated appraisals will contain a larger number of development needs (Section 2).

Table 28 shows the analysis summary for these data. A significant F ratio was found between methods. The coordinated appraisals contained more development needs; hence Hypothesis 12 was confirmed.

The appraisal instructions suggested the supporting facts from the items be used directly in the summary to indicate strengths and needs. However, this procedure was only followed in a few cases by the non-coordinated appraisers. In most instances completely new or much briefer statements were made in these sections. In order to eliminate credit given for merely following instructions these short summaries

TABLE 27

ANALYSIS OF VARIANCE OF THE NUMBER OF PERFORMANCE STRENGTHS, SECTION 1. DATA TRANSFORMED TO SQUARE ROOT OF x + .5

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	11,575	1	11,575	15.58**
Coordinators	72	1	72	
Locations	3,039	1	3,039	4.09*
Interactions:				
Methods x Coordinators	255	1	255	
Methods x Locations	3,069	1	3,069	4.13*
Coordinators x Locations	38	1	38	
Methods x Coordinators x Locations	255	1	255	
Within groups	40,110	54	743	
Total	58,413	61		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

*Significant beyond .05 level
**Significant beyond .01 level

Bartlett's test: Significant at .05 level

TABLE 28

ANALYSIS OF VARIANCE OF THE NUMBER OF DEVELOPMENT NEEDS, SECTION 2. DATA TRANSFORMED TO SQUARE ROOT OF x + .5

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	21,894	1	21,894	55.15**
Coordinators	0	1	0	
Locations	341	1	341	
Interactions:				
Methods x Coordinators	239	1	239	
Methods x Locations	281	1	281	
Coordinators x Locations	179	1	179	
Methods x Coordinators x Locations	248	1	248	
Within groups	20,634	52	397	
Total	43,816	59		

For df 1 and 50: $F_{.05} = 4.03$, $F_{.01} = 7.17$

**Significant beyond .01 level

Bartlett's test: Significant at .05 level

were considered as satisfactory regardless of the adequacy of explanation. Only the thought had to be present. A word count was not made for the same reason. There was no intention to make comparisons based on verbage alone.

A classification of development needs was made using the items and other recurring statements as reference points. For instance, a problem of getting along with others was classed as an Item 9 need even though the item number was not specified. Two additional categories immediately became evident. These were experience and education. Needs which recurred infrequently were attitude, leadership, confidence, responsibility, etc. The latter were all put in the "other" category which was not large enough to break down further. Table 29 shows the number of needs per category. For the coordinated appraisals items,

9 - Ability to work well with others

11 - Observing rules and regulations

12 - Efforts to learn about other phases of the work than his own,

were referred to most frequently. A slight difference was noted between plant and office with the latter emphasizing

17 - Ability to express himself orally

18 - Ability to express himself in writing.

In the non-coordinated group not enough references were made to the items to indicate a need trend.

TABLE 29

COMPARISON BETWEEN COORDINATED AND NON-COORDINATED GROUPS ON CATEGORIES OF DEVELOPMENT NEEDS

Need	Coordinated	Non-coordinated	Total
Items	188	41	229
Experience	31	28	59
Education	14	14	28
Other	17	19	36
Total	250	102	352
$x^2 = 40.817^{**}$			

For 3 df: .05 level $x^2 = 7.815$, .01 level $x^2 = 11.341$
**Significant beyond .01 level

In Table 29 data are shown where a comparison was made between the coordinated and non-coordinated groups in regard to the need categories. A significant chi square indicated a difference between groups. For the coordinated group 76 percent of the needs referred to items. For the non-coordinated group this was forty percent. Eighteen percent of the needs for the coordinated group were in experience and education while this figure was 41 percent for the non-coordinated group.

The methods of handling development needs listed in Section 3 were analyzed according to number, responsibility for action, and type.

Hypothesis 13 - The coordinated appraisals will contain a larger number of methods of handling the development needs.

Table 30 shows the analysis summary. A significant F ratio was found between methods. The coordinated group contained a larger number of methods, thus Hypothesis 13 was confirmed.

The number of strengths, needs and methods of handling the needs was not important in itself. However, a reasonable number of these must be present to assume an adequate coverage of each topic and to furnish a satisfactory amount of material for the feedback interview and subsequent action. There was no evidence of padding or excessive listing of information. The amount of information was usually minimal and the analyses of quantity were within this framework.

The appraisal instructions contained detailed instructions for Section 3. Variety and number of methods, therefore, only indicated a thoughtful consideration of the problem.

Hypothesis 14 - The coordinated appraisals will more frequently indicate action to be initiated by the supervisor or both supervisor and employee, rather than employee.

The appraisal instructions and the coordinators emphasized the point that development action was the responsibility of the supervisor or jointly that of the supervisor and employee. This sort of involvement was considered essential for active interest and participation.

TABLE 30

ANALYSIS OF VARIANCE OF THE
NUMBER OF METHODS OF HANDLING DEVELOPMENT NEEDS

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	413	1	413	16.59**
Coordinators	5	1	5	
Locations	62	1	62	2.49
Interactions:				
Methods x Coordinators	24	1	24	
Methods x Locations	17	1	17	
Coordinators x Locations	2	1	2	
Methods x Coordinators x Locations	5	1	5	
Within groups	1220	49	25	
Total	1748	56		

For df 1 and 50: $F_{.05} = 4.03$, $F_{.01} = 7.17$

**Significant beyond .01 level

Bartlett's test: Not significant

Table 31 gives data where the groups were compared on this factor. The coordinated appraisals placed the responsibility for action on the supervisor or supervisor and employee more frequently than the non-coordinated. Hypothesis

TABLE 31

COMPARISONS BETWEEN COORDINATED AND NON-COORDINATED AND BETWEEN PLANT AND OFFICE IN TERMS OF PERSON PRIMARILY RESPONSIBLE FOR INITIATION OF DEVELOPMENT ACTION

	Coordinated	Non-coordinated	Total
Employee	2	26	28
Supervisor	33	22	55
Both	59	16	75
Total	94	64	158

$x^2 = 39.116$**

	Plant	Office	Total
Employee	14	14	28
Supervisor	28	27	55
Both	31	44	75
Total	73	85	158

$x^2 = 1.333$

For 2 df: .05 level $x^2 = 5.991$, .01 level $x^2 = 9.210$

14 was confirmed. No difference was found between plant and office.

A statement of responsibility for action on the appraisal form did not necessarily mean the supervisor accepted this fact. This was no doubt more true in the coordinated than

the non-coordinated appraisals. It should be a step in the right direction, and in the follow-up by the coordinator the appraisal form served as a basis for fixing the responsibility.

Hypothesis 15 - On-the-job counseling or coaching is suggested more frequently in the coordinated appraisals as a method of training.

Table 32 shows these data. Hypothesis 15 was confirmed. No differences were found between locations.

Previous analyses of development needs at Chrysler have indicated that the majority of the problems are of such a nature that they could best be handled within the department by extra assistance from the supervisor. This is true because of the specific nature of the problems and the generality of classroom training. Therefore, the emphasis has been in this direction through the coordinators and the suggestions for development planning.

It was noted that classroom training received the greatest emphasis in the non-coordinated appraisals even though experience was considered an important need. This method of training suggested in the non-coordinated appraisals seems to contradict experience. Coaching by the supervisor can be the most profitable method of training and meeting the needs in the experience area.

Hypothesis 16 - The development plans of the coordinated appraisals are more specific or concrete than those of the non-coordinated appraisals.

TABLE 32

COMPARISONS BETWEEN COORDINATED AND NON-COORDINATED AND BETWEEN PLANT AND OFFICE IN TERMS OF STATED METHODS OF HANDLING DEVELOPMENT NEEDS

Method	Coordinated	Non-coordinated	Total
Rotation	15	9	24
Coaching, counseling	67	17	84
Classroom course	36	31	67
Self study (reading)	15	5	20
More responsibility	18	6	24
	151	68	219

$x^2 = 12.557$*

Method	Plant	Office	Total
Rotation	12	12	24
Coaching, counseling	46	38	84
Classroom course	24	43	67
Self study (reading)	6	14	20
More responsibility	9	15	24
	97	122	219

$x^2 = 8.741$

For 4 df: .05 level $x^2 = 9.488$, .01 level $x^2 = 13.277$

An important quality aspect of the development plans was labeled specificity or concreteness. It was thought that plans which gave definite direction as to what should be done, by whom, and how were more meaningful than those which were vague and ill-defined. An evaluation of this factor of the development plans was made according to the directions and categories shown in Figure IX. This sorting was done by the coder with a reliability check by the experimenter.

The analysis summary is shown in Table 33. A significant F ratio was found between methods. Hypothesis 16 was confirmed.

Sort of Development Plans

Sort according to specificity - generality or concrete-abstract.
Sort into five stacks according to the categories:

1. Very poor - very unsatisfactory - least
2. Poor - less than expected
3. Fair - average - good as can be expected
4. Good - more than expected
5. Very good - very satisfactory - most

Reliability: r = .69

Figure IX. Directions and Categories for Sorting of Development Plans

TABLE 33

ANALYSIS OF VARIANCE OF SORT OF DEVELOPMENT PLANS ACCORDING TO SPECIFICITY-GENERALITY

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	672	1	672	8.40**
Coordinators	112	1	112	1.40
Locations	253	1	253	3.16
Interactions:				
Methods x Coordinators	52	1	52	
Methods x Locations	36	1	36	
Coordinators x Locations	71	1	71	
Methods x Coordinators x Locations	18	1	18	
Within groups	3916	49	80	
Total	5130	56		

For df 1 and 50: $F_{.05} = 4.03$, $F_{.01} = 7.17$

**Significant beyond .01 level

Bartlett's test: Not significant

Operational Aspects

Two additional factors were investigated which were of practical importance for the administration of such a program but were not directly connected with the information on the appraisal form. These were:

1. The actual time required by the appraiser or appraiser and coordinator to complete the appraisal form
2. The time lapse between the delivery of the forms to the appraiser and the return of the completed forms.

In the non-coordinated group the supervisors were contacted individually following the introductory conferences. At this time they were given opportunity to ask questions about the procedure and appraisal techniques. Each supervisor was given three appraisals, an instruction sheet, and a list of the three employees to be appraised. A time was set as to when the appraiser expected to have the appraisals completed, with the understanding that he would be called at that time and asked to return the forms. This initial period of time was usually a week or ten days.

For the coordinated group the supervisors were contacted for three-hour appointments, at which time three appraisals were conducted. Preliminary experience with coordination using this form indicated that three hours was usually sufficient time to conduct three appraisals.

The coordinators recorded the approximate time spent with each supervisor in the coordinated group. When the

non-coordinated supervisors returned their appraisals they were asked to estimate the time spent in completing the appraisals. In this way it was possible to determine the approximate time spent per appraisal in each case. Summary information is given in Table 34.

TABLE 34

AVERAGE TIME SPENT COMPLETING AN APPRAISAL

	Coordinated	Non-coordinated
Number of supervisors	32	32
Minutes per appraisal (average)	58	113
Range (time in minutes)	30-70	20-360

Table 35 shows an analysis of variance summary for the same data. A significant F ratio was found between methods. A significant Bartlett's test indicated heterogeneity between groups which was anticipated in a comparison of the ranges in Table 34. The important finding here was that the coordinator served as a pacer. He controlled the time involved and insured that the appraisal was given adequate consideration. Twenty minutes was probably not sufficient time to do a good job of appraisal. On the other hand, 360 minutes (6 hours) was no doubt excessive. A good appraisal should be done in considerably less time. This extreme variability in time no doubt was to some extent an important factor in the variation of quantity and quality of appraisals noted earlier.

TABLE 35

ANALYSIS OF VARIANCE OF THE
AVERAGE TIME IN MINUTES PER APPRAISAL FOR EACH APPRAISER

Source of variation	Sum of squares	df	Mean square	F
Between:				
Methods	49,562	1	49,562	13.79**
Coordinators	969	1	969	
Locations	722	1	722	
Interactions:				
Methods x Coordinators	311	1	311	
Methods x Locations	285	1	285	
Coordinators x Locations	4,641	1	4,641	1.29
Methods x Coordinators x Locations	1,859	1	1,859	
Within groups	201,335	56	3,595	
Total	259,684	63		

For df 1 and 55: $F_{.05} = 4.02$, $F_{.01} = 7.12$

**Significant beyond .01 level

Bartlett's test: Significant beyond .01 level

The average time per appraisal was almost twice as much for the non-coordinated group as for the coordinated. In the coordinated situation, of course, two persons were involved rather than one. From the information on time it was possible to get an approximate idea of the relative costs involved. Based on average salaries for supervisors and coordinators it was found that the actual appraisal manpower costs were about:

1) Coordinated - $6.31 per appraisal

2) Non-coordinated - $6.48 per appraisal.

These figures were approximations, but it seemed safe to state that in regard to manpower costs there was essentially no difference between the coordinated and non-coordinated appraisals.

It should be noted that the above costs were based on average salary for first level supervisors. In the management program only supervisors of the second level and above were appraisers. Therefore, the average cost of the supervisor's time was greater and increased with successive levels of supervision. In this case the coordinated appraisals probably cost less per hour than the non-coordinated.

The second factor mentioned is difficult to name. Essentially it involved getting the appraisals finished and returned. In the case of coordination, the coordinator provided the time and place to get the job done. Problems of setting appointments and of cancelled appointments were

encountered. Nine of the 32 coordinated supervisors cancelled their original appointments and had to be re-scheduled. It was necessary to have two sessions with seven supervisors and three sessions with two supervisors because they could not leave their work area for three hours at a time.

On the other hand, 15 of the supervisors in the non-coordinated group of 32 returned their appraisals when called once as scheduled. The period of time here was from four to ten days. Twelve supervisors had to be called from one to seven times more in order to get the appraisals returned. For the non-coordinated supervisors the time required to get the appraisals returned ranged from four to 75 days with a mean of 18 days. This count was based on working days, five per week. Five of the supervisors had the appraisals over seventy days before they returned the forms. It appeared that without coordination the completion of an appraisal program might be extended indefinitely. It must also be kept in mind that only three appraisals were involved in each case.

It was impossible to estimate the cost involved in additional calls, contacts, and delays encountered in the non-coordinated situation. These extra costs do add up, however, and would increase the cost per appraisal as previously estimated.

GENERAL DISCUSSION

The evaluative aspects were those concerned with the ratings alone, the scale values of the judgments and their relationships. In lieu of outside criteria of validity secondary criteria of a rating method were used to compare the two methods of appraisal administration. It was expected that coordination of appraisals would result in improved evaluation of performance. It was thought that this improvement of evaluation would be reflected in the ratings by an increase in discrimination, a decrease in the constant errors of leniency and halo, fewer omitted items and greater reliability for the coordinated appraisals than the non-coordinated.

Failure to confirm the Hypotheses 1 through 8 meant that the coordinated ratings did not differ from the non-coordinated ratings on these criteria. Therefore, an improvement or an increase in the validity of the ratings was unlikely due to the reduction of these errors.

The experimenter suggested that the lack of positive findings appeared to be due to the unexpected characteristics of the non-coordinated ratings. The non-coordinated ratings were more satisfactory on the indexes used than was anticipated. The non-coordinated ratings did not contain the

extreme amount of constant errors usually found in such ratings. Distributions from other plants showed less spread and more central tendency than that found in the non-coordinated ratings. Taylor and Hastman (38), likewise, suggested that their lack of positive findings might be attributed to a situation where the control group gave more desirable ratings than usual.

A more definite emphasis by the coordinators on spread and omissions may have given positive results. The coordinaters were trained and directed to do a careful job of coordination in an effort to duplicate as nearly as possible the operation of other coordinators currently engaged in the ongoing program. This approach was adopted in order to avoid possible criticism that the experimental group or coordinated ratings were atypical or represented an artificial forcing of ratings.

Subsequent observation and experience, not verified experimentally, has indicated differences between coordinators and plants in distributions of ratings. Apparently many situational factors such as acceptance of the program, size of work groups, general management climate, etc. operate to affect the amount of errors present in ratings. The coordinator probably effects more of a change in some situations than others.

In addition, special emphasis by the coordinator can be effective in obtaining more spread. In some cases this may

occur at the expense of good supporting facts. For example, in the present experiment the coordinators could have been directed to insist on a forced distribution of ratings and they likely would have come close to attaining this. Such a distribution for the coordinated ratings in this study would no doubt have given differences regarding spread and skew. Likewise, insistence that no items be omitted would likely have produced differences.

The experimenter has suggested that the effect of the coordinator upon the ratings is probably a function of two factors: (1) the situation, and (2) the emphasis of the coordinator. This study did not support the notion that the coordinator consistently effects a change in the ratings on the characteristics used as indexes.

An improvement in the ratings for the coordinated appraisals in terms of the criteria used was not supported by the findings relevant to Hypotheses 1 through 8.

The criteria of the appraisal as a development tool were established from an examination of its function in the development process. It was necessary to derive the criteria in this fashion because the specific role of the appraisal was considered important and because there was no previous research on development devices which appeared relevant to this system. The steps in the development procedure are:

1. Performance appraisal of the employee by his immediate supervisor

2. Appraisal review by the appraiser's immediate superior

3. Performance interview with the subordinate by his supervisor

4. Training and development of the subordinate

5. Follow-up on the course of action by the coordinater

Hypotheses 9 through 16 focused attention on the development procedure itself.

The supporting facts and the performance summary are both used in the review to acquaint the reviewer with the status of present performance and the development plans. The performance summary serves as the basis for the interview, training and development, and follow-up. The supporting facts are not used in these latter phases of the process. In order to fully utilize these steps adequate information should be contained in the appraisal form.

According to the indexes used the supporting facts of the coordinated appraisals were greater in quantity and were more specific or factual than those in the non-coordinated appraisals. In addition to the indexes developed for quantity and specificity the improvement in supporting facts was corroborated by an expert evaluation of them plus an estimation of ratings based on the information contained in the facts. The coordinator was instrumental in replacing generalized comments about performance with specific examples or concrete instances of behavior which provided more

meaningful information for the review and the writing of the performance summary.

The general attitude, as reflected by positive or negative comments, of the appraiser toward the ratee was not changed by the presence of the coordinator. His criticism of the employee did not become more or less severe in the coordinated situation. This trend was considered as a favorable sign which suggested that the coordinator was accepted by the appraiser and that the coordinator did not exert undue pressure in questioning the appraiser.

Although the appraisers did not differ in their use of so-called dogmatic words, the coordinated appraisers appeared to be more cautious in their statements by the use of more qualifying words. The latter finding suggests that the appraiser considered more carefully his statements which were recorded as supporting evidence on the coordinated appraisals. Employee performance is probably better characterized by such words as usually, frequently, etc. than by words as never, none, etc.

Undiplomatic probing or extreme insistence by the coordinator could result in a defensiveness on the part of the appraiser which would slant the information obtained. It did not appear that the improvement in supporting facts was obtained at the expense of distorting due to this factor.

The results obtained in respect to the number of no responses in the performance summary were among the most

definitive findings of this study. A satisfactory interview with specific suggestions for development, definite development action, and a detailed follow-up can not be expected unless information is available to give direction to the process. The coordinated appraisals were more satisfactory in regard to completion of the form than the non-coordinated. In the latter group a significantly greater number of sections were completely omitted. It appears that a coordinator is essential to insure that material is recorded at this point in the process for use later.

The number of strengths, needs and methods of development recorded in the performance summary in itself was not considered important. Too much material in the summary could lead to confusion and be as detrimental as a lack of information. An over-abundance of information was not a problem. The average number of strengths, needs and methods for the non-coordinated group was two, one and one. For the coordinated group these numbers were three, three and two, respectively. The coordinated appraisals did contain a greater number of strengths, needs and methods than the non-coordinated appraisals. These findings indicate that a more adequate amount of information for the performance interview and subsequent development action was obtained in the coordinated appraisals.

The analysis of development needs indicated that the needs were more frequently related to the items for the

coordinated appraisals than the non-coordinated. In the non-coordinated appraisals needs in the areas of experience and education were more frequent than in the coordinated. These results would suggest that the summary of needs was more truly a summary of the item facts in the coordinated appraisals than in the non-coordinated. The item facts are more definitive of the needs because of their specificity and individuality. The needs of experience and education represent rather general needs which might apply to everyone.

The responsibility for the training and development of employees is considered a line management function. The personnel staff assists line management with techniques and people in carrying out this function. Each supervisor has the responsibility for the training and development of his subordinates. Furthermore, experience has indicated that the majority of development needs can be handled most satisfactorily in the department by the supervisor.

An aim of the appraisal program was to begin the education of supervision in terms of the above principles. Efforts were made to aid supervision in the recognition of this responsibility for training and development. The coordinators were directed to emphasize the appraiser's role in the entire process. The appraisal instructions were also slanted in this direction.

The findings indicated that the coordinated appraisals placed the responsibility for action on the supervisor or

supervisor and employee more frequently than the non-coordinated appraisals. Coaching or on-the-job counseling was suggested more frequently by the coordinated appraisers than the non-coordinated. The recording of these approaches on the appraisal form does not mean acceptance by the supervisor but it does indicate what will be expected from him through follow-up.

A significant F ratio between coordinators was found on the three factors: minor thought count, number of criteria, and expert evaluation. The appraisals conducted by coordinator B contained more minor thoughts and criteria and were valued higher by the experts than the appraisals of coordinator A. Differences between coordinators can be expected. When they become too large, however, this may be indicative of a training need for coordinators.

No significant differences were found between locations. Coordination was equally effective in the office and plant. It might be expected that the office supervisors would benefit less from coordination than the plant supervisors because their work is primarily paperwork. No differences of this kind were observed.

Although heterogeneity of variance reduced the precision of the F test in several instances the information gained from a significant Bartlett's test was also useful. Coordination produced increased uniformity between appraisers on quantity and quality of supporting facts. Likewise the

extreme variation in time spent per appraisal was reduced by the coordinator.

Evidence has been produced which indicated that the coordinator did effect an improvement in the appraisal as a development tool. In fact, if the appraisal is to fulfill its intended function in the development process the coordinator is essential.

The procedure used here for the evaluation of the appraisal as a development tool was not new. The elements of the procedure were found in other types of research. A systematic and thorough application of these techniques to a problem such as this was not found in the literature. The emphasis in management circles today is on development, although little has been done in the way of research because of the magnitude of the problem. The experimenter suggests that a solution might be a series of studies at various steps in the program.

The next important phase for study in the Chrysler Management Development is the performance interview. Individual development depends upon the communication of the appraisal information to the ratee and his acceptance of the development plans.

SUMMARY AND CONCLUSIONS

Sixty-four first-line supervisors each rated or appraised the job performance of three subordinates. Two methods of administration, two coordinators, and two locations identified as coordinated vs. non-coordinated, coordinators A and B, and plant vs. office, respectively, were incorporated into a factorial design. Method of administration, with or without a coordinator, was the variable of primary interest. The coordinator was a staff person who assisted in appraisal by questioning the appraiser and recording the information in a modified Field Review type interview.

The objectives of appraisal were: (1) evaluation of present performance and (2) planning for individual improvement. These goals were considered separately in the experiment. It was predicted that the coordinated appraisals would be superior to the non-coordinated in meeting these aims.

The relative merits of the methods as a system of evaluation were inferred by comparing the treatments on secondary criteria or characteristics of a rating method. Satisfactory outside criteria were not available. It was predicted that the ratings of the coordinated appraisals would be improved by increased discrimination, reduced leniency, reduced halo, increased coverage and increased comparability between ratings. Specifically, the coordinated

appraisals would differ from the non-coordinated in the following respects. The coordinated ratings would show:

1. Increased spread of overall ratings
2. Increased variability of ratings for an appraiser
3. Mean of ratings nearer central scale value
4. More symmetrical distribution of ratings
5. Greater variability of item ratings
6. Smaller proportion of identical ratings
7. Fewer omitted responses
8. Greater relationship to separately obtained ranks.

The above statements were not supported by the findings. No differences were found between the methods, coordinators or locations on these factors. The conclusion was that the coordinator did not improve the effectiveness of the appraisal as a rating instrument. The experimenter offered additional evidence which suggested that the negative findings might be due to the fact that the non-coordinated ratings were improved on these characteristics over what was usually found.

Immediate criteria used to evaluate the appraisal as a development instrument were established from an examination of its function in the complete development procedure. Subsequent to the appraisal session the information recorded on the form was to be used for (1) a review by the appraiser's supervisor, (2) a performance interview with the employee, (3) training and development of the subordinate, and (4)

follow-up on action by the coordinator. A content analysis was conducted on the response material found in certain sections of the appraisal to determine whether the coordinated appraisals exceeded the non-coordinated on quantity and quality of information. The areas of the form important to the above steps were the supporting facts and performance summary which contained sections for a summary of individual performance strengths, development needs and development plans.

The findings indicated that the supporting facts of the coordinated appraisals:

1. Contained a greater amount of information
2. Were more descriptive of specific performance
3. Contained more criteria of performance
4. Contained more examples or illustrations of specific instances
5. Contained more qualifying words

than the non-coordinated appraisals. The supporting facts of coordinator B were superior to those of coordinator A on three factors which indicated variation between coordinators. There was no difference between the coordinated and non-coordinated supporting facts in terms of number of negative and positive comments or the use of dogmatic words.

The non-coordinated group of appraisals contained a greater number of appraisals with Sections 1, 2, and 3 of the performance summary omitted.

The performance summaries of the coordinated appraisals contained a larger number of the items listed below than the non-coordinated appraisals.

1. Performance strengths
2. Development needs
3. Methods of handling development needs.

The development needs for the coordinated group were more closely related to the supporting facts. The coordinated appraisals placed the responsibility for the initiation of development on the supervisor or both supervisor and employee rather than the employee only and suggested on-the-job counseling or coaching as a method of development. The development plans for the coordinated appraisals were more specific than those for the non-coordinated. There were no differences between coordinators in terms of the performance summary. No differences were found between the plant and office on any of the development criteria.

It was concluded that the coordinator did effect an improvement on the appraisals as a tool for development in terms of the outlined procedure and criteria used for evaluation. In fact it can be said that the coordinator plays an essential role and makes a significant contribution to the development procedure.

APPENDIX

Coordinator A

Coordinated Function	Group size	Non-coordinated Function	Group size
		Plant	
Janitor	29	Janitor	14
Milling Machine	15	Milling Machine	35
Inspection	9	Inspection	4
Tool Room	13	Heat Treat	7
Heat Treat	5	Steamfitter	30
Electrician	17	Tool Cribs	16
Tool Stores	5	Boilerhouse	13
Sheet Metal Equipment	6	Machining	16
		Office	
Sub-assembly	4	Laboratory	4
Tool Engineering	13	Tool Engineering	15
Tool Engineering	11	Tool Engineering	8
Time Study	7	Time Study	12
Accounting	6	Telephone Operators	9
Comptometer	5	Laboratory	5
Laboratory	4	Cost Estimate	11
Budget	3	Machine and Tool Layout	13

Figure X. Listing of Departments in the Sample by Function and Group Size

Coordinator B

Coordinated Function	Group size	Non-coordinated Function	Group size
		Plant	
Milling Machine	22	Milling Machine	5
Planning	6	Planning	32
Inspection	6	Inspection	8
Tool Room	36	Tool Room	17
Heat Treat	4	Heat Treat	3
Millwrights	15	Machine Repair	19
Sheet Metal and Machine Assembly	14	Sheet Metal and Machine Assembly	25
Inspection Laboratory	4	Machine Laboratory	3
		Office	
Machine and Tool Layout	14	Planning	8
Planning	3	Tool Engineering	4
Plant Engineering	13	Plant Engineering	5
Accounting	6	Accounting	7
Accounting	5	Accounting	4
Master Mechanics	7	Master Mechanics	8
Resident Engineering	8	Engine Test	7
Tabulating	8	Traffic	3

Figure X. Continued

PERFORMANCE REVIEW

NAME_____ DATE_____

DEPARTMENT_____ PLANT_____

JOB ASSIGNMENT_____

CLASSIFICATION NO._____ TIME ON JOB_____

CHRYSLER CORPORATION

PERSONNEL DEVELOPMENT

Figure XI. Appraisal Instrument

PART I

On each of the following items, consider how the person being rated performs these functions in terms of what you expect. For each item give specific incidents which illustrate the person's performance. After each item check one of the five choices on the scale which best describes his standing. The supporting facts should form the basis for this decision. If there has been no chance to observe the ratee on an item, place a "DK" for "Don't Know" after the item.

CHARACTERISTICS OF WORK-OUTPUT: Consider only the quality and quantity of work performed.

Supporting Facts

1. What is the quality of work done? _____ 1 2 3 4 5

2. How much work does he turn out? _____ 1 2 3 4 5

PERFORMANCE EVALUATION: Consider how the person actually does the work.

3. Effectiveness in working under little supervision. _____ 1 2 3 4 5

4. Dependability in carrying out assigned tasks. _____ 1 2 3 4 5

5. Suggesting improvements in ways of doing work. _____ 1 2 3 4 5

6. Handling non-routine work situations effectively. _____ 1 2 3 4 5

7. Getting work out on time. _____ 1 2 3 4 5

8. Assuming added duties in order to get the job done. _____ 1 2 3 4 5

9.	Ability to work well with others.	1	2	3	4	5
10.	Thinking a problem through before acting.	1	2	3	4	5
11.	Observing rules and regulations.	1	2	3	4	5
12.	Efforts to learn about other phases of the work than his own.	1	2	3	4	5
13.	Following job instructions.	1	2	3	4	5
14.	Speed in learning new duties.	1	2	3	4	5
15.	Keeping superiors informed on matters of mutual concern.	1	2	3	4	5
16.	Accepting without complaint, adverse work conditions and difficult job assignments.	1	2	3	4	5
17.	Ability to express himself orally.	1	2	3	4	5
18.	Ability to express himself in writing.	1	2	3	4	5

CODE: 1 = Unsatisfactory
2 = Less than expected
3 = Expected
4 = More than expected
5 = Outstanding

PART II

SUMMARY OF PERFORMANCE EVALUATION

1. What are his significant strengths and abilities as demonstrated on the job?

2. What are his significant developmental needs for more effective performance?

3. Recommended methods of development:

OVER-ALL RATING

4. Considering all factors, how would you rate the over-all performance of this person?

( ) Unsatisfactory ( ) Expected ( ) More than expected

( ) Less than expected ( ) Outstanding

PART III

EVALUATION OF FUTURE GROWTH

Future growth potential is the possession of the ability to assume and perform managerial responsibilities. This is determined by the employee's present and past performance plus a consideration of the following factors:

a. Age, health and stamina.

b. Desire to accept higher responsibilities.

c. Present status in the organization.

d. The possession of leadership qualities

In regard to the above, this person is:

( ) Probably not managerial material.

( ) Merits consideration for advancement to a managerial position with further development. (How long before ready? _____)

( ) Merits consideration now for advancement to a managerial position.

Indicate any other type of work for which the person is better suited. (Specify)

Additional Comments:

Appraiser_____ Title_____ Date_____

Figure XII.

INSTRUCTIONS FOR APPRAISAL

These instructions are to assist you in completing the Performance Review form.

PART I -- The rating is made on a 5-point scale which is illustrated at the bottom of the page. The scale categories can be defined as follows:

1. Unsatisfactory - Consistently fails to meet minimum requirements of the job.
2. Less than expected - Performance somewhat below normal job requirements.
3. Expected - Performance satisfactory. Meets normal requirements of the job.
4. More than expected - Performance somewhat exceeds normal requirements of the job.
5. Outstanding - Performance exceptionally good. Far beyond usual job requirements.

Appraise 3 or 4 of your people at the same time. Rate all of them on item 1, then go on to item 2, etc.

For each item under "Supporting Facts":

1. Give specific examples to illustrate how the person does this part of his job.
2. The supporting facts should be things which happen often rather than once in a while.
3. Do Not merely restate the item as supporting facts.

Assign a rating 1 - 5 based on the facts you have given. Keep in mind that: a) employees vary in performance and, to this extent, should be rated differently; b) it is only the unusual person who is either high, average, or low on all aspects of job performance.

PART II -- Summarize your evaluation and list development plans.

For questions 1 and 2: List the item numbers (from Part I), together with supporting facts, which show:

Question 1. The employee's strengths or where he does best.
Question 2. His developmental needs or where he can improve.

You may want to qualify an item or add an area not covered by the appraisal, but in all cases be sure to give supporting facts. Discuss job performance not personality traits.

Under statement 3, list suggested plans for individual development directed toward meeting the needs outlined in Question 2, plus other action which may help prepare the employee for advancement. The plans may include: a) steps your employee might take to improve.
b) steps you might take to help him improve.

The plans should help your employee in either or both of two ways:

1. Help him do better on his present job, or
2. Prepare him for advancement.

Your suggestions will be easier to follow if you:

1. Are specific - say exactly what is to be done.
2. Are realistic - suggest things a person is able to do and can achieve within a year.
3. List all of your suggested plans.

Your recommended action may be in one or more of these fields:

1. Things on which you can advise and guide him.
2. Training he can get in the department.
3. Special problems or other jobs to give experience.
4. Recommended reading or outside courses (college, night school, Chrysler Institute).

Question 4 asks you to make a single over-all rating.

Consider: a) the supporting facts and ratings,
b) give more weight to the most important items.

PART III -- Indicate whether or not this person should be considered for advancement into management with reasons why. When giving these reasons, be sure to comment on factors a, b, c, and d.

BIBLIOGRAPHY

1. Balch, D. E. Executive selection, development, and inventory. Personnel Series No. 171. New York: Amer. Mgmt. Ass., 1957.

2. Bellows, R. M. Psychology of Personnel in Business and Industry (2nd ed.). New York: Prentice-Hall, 1954.

3. Benjamin, R., Jr. A survey of 130 merit rating plans. Personnel, 1952, 29, 289-292.

4. Bingham, W. V. D. Halo, invalid and valid. J. appl. Psychol., 1939, 23, 221-228.

5. Chi, Pan-Lin. Statistical analysis of personality rating. J. exp. Educ., 1937, 5, 229-245.

6. Dooher, M. J. and Marquis, Vivienne (Eds.). Rating Employee and Supervisor Performance: A Manual of Merit Rating Techniques. New York: Amer. Mgmt. Ass., 1950.

7. Driver, R. S. The validity and reliability of ratings. Personnel, 1941, 17, 185-191.

8. Edwards, A. L. Experimental Design in Psychological Research. New York: Rinehart, 1950.

9. Ferguson, L. W. The development of a method of appraisal. Personnel, 1947, 24, 127-136.

10. Flanagan, J. C. A new approach to evaluating personnel. Personnel, 1949-50, 26, 35-42.

11. Flanagan, J. C. & Burns, R. K. The employee performance record: A new appraisal and development tool. Harvard Business Review, 1955, 33(5), 95-102.

12. Freyd, M. The graphic rating scale. J. educ. Psychol., 1923, 14, 83-102.

13. Grant, D. L. A factor analysis of managers' ratings. J. appl. Psychol., 1955, 39, 283-286.

14. Guilford, J. P. Psychometric Methods (2nd ed.). New York: McGraw-Hill, 1954.

15. Hoel, P. G. Introduction to Mathematical Statistics. New York: John Wiley & Sons, 1947.

16. Institute for Social Research, Survey Research Center. A Manual for Coders. Ann Arbor: University of Michigan, 1955.

17. Johnson, D. M. The Psychology of Thought and Judgment. New York: Harper & Brothers, 1955.

18. Johnson, D. M. & Vidulich, R. N. Experimental manipulation of the halo effect. J. appl. Psychol., 1956, 40, 130-134.

19. Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. New York: Houghton Mifflin, 1953.

20. Lynch, J. M. The psychology of the rating scale. Educ. Admin. & Supervis., 1944, 30, 497-501.

21. McMurry, R. N. & Co. Patterned Merit Review Plan (Manual). Chicago: Dartnell Corp., 1950.

22. Mahler, W. R. Let's get more scientific in rating employees. Personnel, 1947, 23, 310-320.

23. Mahler, W. R. Twenty Years of Merit Rating. New York: Psychol. Corp., 1947.

24. Mahler, W. R. An experimental study of two methods of rating employees. Personnel, 1948, 25, 211-220.

25. Patterson, C. H. On the problem of the criterion in prediction studies. J. consult. Psychol., 1946, 10, 277-280.

26. Poffenberger, A. T. Principles of Applied Psychology. New York: Appleton-Century, 1942.

27. Shaeffer, R. E. Merit rating as a management tool. Harvard Business Review, 1948, 27, 693-705.

28. Sission, D. E. Forced choice - the new army rating. Personnel Psychol., 1948, 1, 365-381.

29. Smyth, R. C. & Murphy, M. J. Job Evaluation and Employee Rating. New York: McGraw-Hill, 1946.

30. Spicer, L. G. A survey of merit rating programs in industry. Personnel, 1951, 27, 515-518.

31. Springer, Doris. Ratings of candidates for promotion by co-workers and supervisors. J. appl. Psychol., 1953, 37, 347-351.

32. Standard Oil Company, Employee Relations Department. Made to Measure. New Jersey: Author, 1951.

33. Stevens, S. N. & Wonderlic, E. F. An effective revision of the rating technique. Personnel J., 1934, 13, 125-134.

34. Stockford, L. & Bissel, H. W. Factors involved in establishing a merit-rating scale. Personnel, 1949, 26, 94-116.

35. Symonds, P. M. Diagnosing Personality and Conduct. New York: Appleton-Century, 1931.

36. Taylor, E. K. & Manson, Grace E. Supervised ratings - Making graphic scales work. Personnel, 1951, 27, 504-514.

37. Taylor, E. K., Schneider, Dorothy E., & Symons, Nancy A. A short forced-choice evaluation form for salesmen. Personnel Psychol., 1953, 6, 393-401.

38. Taylor, E. K. & Hastman, R. Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychol., 1956, 9, 181-206.

39. Thorndike, E. L. A constant error in psychological ratings. J. appl. Psychol., 1920, 4, 25-29.

40. Tiffin, J. Industrial Psychology (3rd ed.). New York: Prentice-Hall, 1952.

41. Wadsworth, G. W. Jr. The field review method of employee valuation and internal placement, I - VI. Personnel J., 1948-49, 27, 47-54, 99-106, 135-141, 183-190, 227-232, 263-268.

42. Whitla, D. K. & Tirrell, J. E. The validity of ratings of several levels of supervisors. Personnel Psychol., 1953, 6, 461-466.

ROOM USE ONLY

[illegible] dept.

[illegible]

JUL 11 1960

ROOM USE ONLY

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03056 7097