To the Graduate Council:

I am submitting herewith a dissertation written by Robert Earl Burt entitled "The Equivalence of Job Analysis Agents: A Methodological Reappraisal." I recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial and Organizational Psychology.

Michael E. Gordon, Major Professor

We have read this dissertation
and recommend its acceptance:

Accepted for the Council:

Vice Chancellor
Graduate Studies and Research

THE EQUIVALENCE OF JOB ANALYSIS AGENTS: A METHODOLOGICAL REAPPRAISAL

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Robert Earl Burt

December 1980

DEDICATED

TO MY PARENTS

ACKNOWLEDGMENTS

The opportunity to formally acknowledge the guidance and direction provided by my major professor, Dr. Michael Gordon, is a pleasure. His critical comments at each stage in the preparation of the dissertation combined with his constant comradery created a supportive and stimulating atmosphere. It has indeed been my honor to work with this man whose dedication to the discipline is unsurpassed.

I wish to express my appreciation for the help received from the other members of my doctoral committee. Dr. John Philpot served as my enthusiastic and most knowledgeable guide into the fascinating world of multivariate statistics. From Dr. Eric Sundstrom, I received my first "hands-on" experience in conducting research and from whom I continue to learn. Dr. Jack Larsen provided encouragement, suggestions, and very important assistance in locating an appropriate source of data.

Also deserving special mention is my friend and colleague, Kammer Boyle. Her insightful comments have contributed invaluably to both my personal and professional development throughout graduate school.

The dissertation could not have been completed without the cooperation of individuals from PAQ Services, Inc., and Personnel Decisions Research Institute (PDRI). Dr. Robert Mecham, Director of PAQ Data Processing Division, informed me of the job analysis project which PDRI had completed, and later provided necessary computer processing of the raw PAQ ratings at a much appreciated research discount. Dr. David Bownas, then with PDRI, was extremely helpful in making available for

my dissertation actual job analysis information collected from job incumbents and their supervisors. I am very grateful for the time invested and concern expressed by Dr. Mecham, Dr. Bownas, and the director of PDRI, Dr. Marvin Dunnette.

Financial assistance from the Walter Melville Bonham Dissertation Award presented by the Graduate Fellowship Awards Committee of the College of Business Administration allowed me to devote a major portion of my time to completion of the research reported in this dissertation.

ABSTRACT

The purpose of this research was to conceptually and empirically investigate the equivalence of ratings provided by different types of job analysis agents. Specifically, ratings of the job of power-plant operator given by job incumbents, their supervisors, students with access to the job description, and students who were only told the job title were examined for equivalence.

Several existing methods for assessing similarity of ratings were reviewed before the empirical investigation was undertaken. Because the correlational approaches appeared deficient for this purpose, a definition of rater equivalence was developed which involves the use of distance measurements in multivariate space. By this definition, the ratings from different types of raters are considered equivalent if and only if the centroids and the dispersions of their multivariate distributions are the same. Equivalence of centroids assures that significant differences do not exist between types of raters in their mean ratings of a job on different dimensions. Equivalence of dispersions (measured by median squared distance from the group centroid) guarantees that interrater agreements (within groups of different types of analysts) do not differ significantly.

Four types of raters completed the Position Analysis Questionnaire (PAQ). One power-plant operator and his or her supervisor from each of 94 plants in 30 companies across the country rated the incumbent's job. Undergraduate management students also completed the PAQ for the job of

power-plant operator. Half were provided with a description of the job (n = 94) and half were told only the job title (n = 94). The dependent variables were the 13 overall factor scores obtained from the individuals' responses to the PAQ items.

Both the correlational approach for assessing similarity and the set of multivariate techniques developed from the definition of rater equivalence were applied to the same set of sample data. Significant correlations among different types of raters over the 13 PAQ dimensions suggested that the source of job analysis information makes little difference. However, an examination of the equivalence of dispersions and the equivalence of centroids produced very different results than that provided by the correlational analyses concerning the equivalence of different job analysis agents. Significantly more agreement (measured by means of dispersion) existed among incumbents than among either group of students. Supervisors displayed greater interrater agreement than did students given only the job title. Moreover, the distance between each pair of group centroids was statistically significant (measured by means of multivariate F ratio). No two groups of analysts met the requirements for their ratings to be considered equivalent.

The findings have implications for procedures used in collecting job analysis information and for methods used to assess the equivalence of raters. The sensitivity of the PAQ was demonstrated by the very large differences in ratings between those provided by students and those given by individuals more familiar with the job (viz., incumbents and their supervisors). Also, the findings suggested that practitioners should expect modest differences in job analysis information depending

on whether the information is provided by incumbents or by their supervisors. Perhaps most importantly, the results further demonstrated that the choice of methods employed to determine the equivalence or similarity of different types of raters can substantially affect a researcher's conclusions. The definition of equivalence which was developed and applied to actual data is offered as an improvement over the more traditional correlational approaches.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Possibly more than any other topic, job analysis is central to the discipline of industrial and organizational psychology. Uses of job analysis information include organizational design, personnel administration, and work and equipment design (Cascio, 1978, p. 132). In response to governmental pressure for empirical evidence of test of validity, industrial psychologists have invested considerable energy in developing sophisticated methods for analyzing job analysis data in order to determine the extent to which a group of jobs is similar or different, i.e., the formation of job families (cf. Dunnette & Borman, 1979).

However, no matter how important or exotic applications of job analysis become, the adequacy of a job analysis rests ultimately on the adequacy of the information provided by people about the job. Three types of people have sufficient knowledge of a job to describe it in a job analysis—job incumbents, their supervisors, and job analysts (McCormick, 1976, p. 654). A pragmatic question of much importance to researchers and practitioners using job analysis data is whether the different agents provide equivalent job information. Can they be used interchangeably when conducting a job analysis?

During the past 50 years, a number of researchers have attempted to determine if people with different relations to the job provide different job information. Recently, Smith and Hakel (1979) argued that students presented only with job titles or job specifications may be used as sources of job information, as well as the three types of job analysis

agents specified by McCormick (1976). Moreover, "there is little difference between analyst sources, including students, in terms of their ability to reliably analyze jobs" (p. 677), and students can provide "response profiles that are indistinguishable from those given by incumbents, supervisors, and job analysts (when similarity of mean profiles is measured with the Pearson correlation coefficient)" (p. 685). This implies that the extent of one's knowledge about a job does not affect one's rating of the job, at least when using the structured job analysis questionnaire administered in their study. This conclusion casts serious doubts on the meaningfulness of job analysis information derived from such an instrument. The statement is most disconcerting when one realizes that the job analysis instrument in question is one of the most carefully developed and best researched structured job analysis questionnaires in existence—the Position Analysis Questionnaire (PAQ; McCormick, Jeanneret, & Mecham, 1972).

However, the issue of the equivalence of different job analysis agents remains open. Methodological problems inherent in the data analysis performed by Smith and Hakel (1979) prevent a clear interpretation of their findings. A careful review shows that the article by Smith and Hakel (1979) is only the most recent addition to a body of literature replete with methodological difficulties. Conclusions about this issue drawn by previous researchers need be reconsidered in terms of the limitations imposed by the design of their studies. No definitive statement concerning rater equivalence is justified on the basis of the empirical investigations appearing in the literature. In the present research, the equivalence of ratings from different job analysis agents

is investigated by logically developing a new definition of rater equivalence from a multivariate perspective, challenging the conclusions of relevant past research in light of the respective methodological weaknesses, and providing empirical evidence on this research question by means of methods consistent with the newly developed definition of equivalence.

This dissertation is organized into six chapters. Chapter I is devoted to developing and defending a definition of equivalence. In Chapter II, methods used to assess interrater agreement through profile analysis are critiqued so that the review of the literature in which interrater agreements are assessed presented in Chapter III may be more cogent. The research method is discussed in Chapter IV. In Chapter V, the statistical procedures which were used in assessing the equivalence of job analysis agents are developed and the obtained results are presented. Chapter VI consists of discussion and conclusions.

CHAPTER I

DEFINITION OF EQUIVALENCE

When viewed in multidimensional space, the ratings of a job on several dimensions from a single rater can be represented by a single point. The coordinates of this point are the ratings assigned to each of the job dimensions.

The sets of ratings provided by several agents of the same type (i.e., job analyst, incumbent, or supervisor) create a distribution of points in the multidimensional space which may be visualized as a cloud. This cloud of points has two salient characteristics: centroid and dispersion. The centroid is defined in terms of the mean rating on each of the job dimensions. Dispersion refers to how compact or spread-out the cloud is. Both characteristics must be considered in developing a definition for the equivalence of job analysis agents. The ratings from different types of raters are considered equivalent if and only if both the centroids and the dispersions of their multivariate distributions are the same. Pragmatically, this definition requires that neither the centroids nor the dispersions of the distributions differ significantly.

I. CENTROIDS

Centroids which do not differ signify that the average ratings of a job from different agents are the same across all of the job dimensions. The centroids—or "centers" of the clouds—are located in the same area of the multidimensional space.

The importance of this requirement is accentuated when considering job analysis as the basis for the formation of job families. Determining which jobs are similar and which are different has recently received much attention because of its importance for validity generalization, wage and salary administration, and performance appraisal systems (Cornelius, Carron, & Collins, 1979). Some methods of identifying job families are based on the position of the jobs in multidimensional space as determined by their ratings on several job dimensions (e.g., Taylor, 1978; Tornow & Pinto, 1976). In such procedures, jobs which cluster together (i.e., are close together in the multidimensional space) are categorized as belonging to the same job family. If the centroids of the ratings provided by different job analysis agents are different , the position of the job in space is dependent on who does the rating. Conceivably, a job might be classified as being a member of job family A if rated by an incumbent and as belonging to job family B when a supervisor performs the analysis. Ratings from different agents would clearly not be interchangeable in such a situation, and different agents could not be considered equivalent.

Pursuing this idea, consider the definition of job equivalence provided by Lissitz, Mendoza, Huberty, and Markos (1979, p. 521):

> In order for two jobs to be equivalent (i.e., not different), the means of the rater's scores on all the dimensions must not be significantly different. From a statistical point of view, the question being asked is whether the centroids, defined by the mean dimension scores, are significantly different.

Suppose that a single job is rated by several incumbents and several job analysts, and an error is made. The ratings from the two groups of job

analysis agents are accidentally coded as belonging to two different jobs. The Lissitz et al. (1979) definition of the equivalence of jobs is invoked and the centroids of the two "jobs" are found to be significantly different. The researcher concludes that the two "jobs" are not equivalent. Later the mistake is discovered and the researcher is forced to amend his conclusion to read: The job is not equivalent to itself. For different types of raters to be considered equivalent, their descriptions of a job should not lead to such seemingly paradoxical results.

II. DISPERSIONS

Dispersion refers to how compact or spread out the sets of ratings from different individuals are about the mean ratings of the group (i.e., centroid). Conceptually, dispersion refers to the size of the cloud of points in multidimensional space. If the dispersion is very small, it means that the group of judges rated the job very similarly across all the job dimensions. Conversely, a large dispersion signifies much disagreement among the members of the group in their ratings of the job.

Even if the centroids of the ratings from different job analysis agents are the same, an examination of the dispersions of the ratings must be undertaken in order to determine the equivalence of the different agents. Equal centroids and equal dispersions are required for equivalence.

Dispersion is defined here as the average squared distance of the points from the centroid in standardized space. This may be operationalized by using the Mahalanobis d^2 (Tatsuoka, 1971) so that the

distance measured in any direction is in terms of comparable units. (A more detailed discussion of the Mahalanobis $\underline{d}^2$ is presented in Chapter II when considering methods of assessing profile similarity.)

$\underline{d^2 \text{ versus } \Sigma}$. Another definition of dispersion is based upon the within-group variance-covariance matrix (denoted Σ). In fact, Tatsuoka (1971, p. 66) used the terms "variance-covariance matrix" and "dispersion matrix" synonymously. Both forms refer to a multivariate analogue of variance, a sort of generalized variance. Nevertheless, comparing the dispersions of two groups is different than testing for the homogeneity of the within-group variance-covariance matrices.

Defining equivalence in terms of dispersion is less stringent than requiring that $\Sigma_1 = \Sigma_2$. This is illustrated in Figure 1. The two ellipses represent the scatterplot of points in the two-dimensional space from Group 1 and Group 2. Assume that the marginal distributions on x_1 are the same for both groups, and that they are also the same on x_2. Although the variances of the two groups are assumed equal, the covariance between x_1 and x_2 is positive in one group and negative in the other. If the homogeneity of the within-group variance-covariance matrices was tested ($H_o: \Sigma_1 = \Sigma_2$) with a large enough sample size, a significant difference would be found due to the different covariances.

Examination of dispersions produces different results. The average squared distance from the centroid would be the same in both groups. The sizes of the two "clouds" are the same and the dispersions are also equal.

In general, a difference in dispersions indicates heterogeneity of the within-group variance-covariance matrices. However, as illustrated

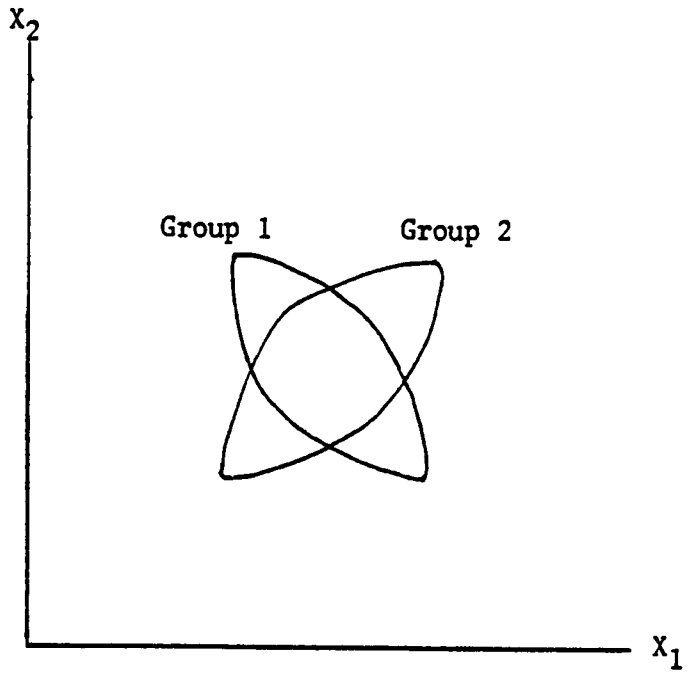


Figure 1. Dispersions ($\overline{d}^2$) are the same although $\Sigma_1 \neq \Sigma_2$.

in Figure 1, $\Sigma_1 \neq \Sigma_2$ does not imply the groups differ in dispersions. Nevertheless, the two conceptions of generalized variance are very similar. Dispersion is preferred because of its simplicity, applicability in small samples, and its relation to "error" in classical test theory. Before discussing related statistical and pragmatic reasons why dispersions must be considered in determining the equivalence of job analysis agents, the relationship with test theory is explored.

Disagreement as error. When several judges rate the same job, disagreement in their ratings is measured by dispersions. Two sources for the dispersion may be identified—actual differences in job content and rating errors. The former is examined first.

Two incumbents of a job (e.g., secretary) may perform somewhat different duties. Their ratings of the job reflect this actual difference in what they do. Similarly, this difference may also be present when supervisors rate their subordinates' jobs, and job analysts may possess somewhat different information about what occurs in a particular job category. Carried to the extreme, no two people perform exactly the same sort of activities even though they may have the same job title.

When using job analysis information, one is usually interested in describing the "archetypal" job—the job as it would be if it were identical for every job incumbent. Minor deviations from the archetype represent idiosyncratic differences in the job content for different incumbents. In trying to describe the archetypal job such deviations may be considered as error even if the ratings of the actual job content are perfectly accurate.

A second source of dispersion stems from the inability of raters to analyze a job perfectly. The existence of such error whenever humans judge something is a tenet of industrial and organizational psychology. The analysis of rater errors has been undertaken by a host of researchers. For example, Cascio (1978, pp. 320-322) reviewed some of the better known constant errors in rating (viz., leniency, central tendency, halo effect, proximity, and logical errors). Other authors have presented similar discussions (e.g., Landy & Trumbo, 1976; McCormick & Tiffin, 1974) and the evidence is overwhelming that rating errors do occur.

Dispersion has been shown to be caused by idiosyncratic differences in job content and human fallibility, both of which may be treated as error. Although the classical test theory concept of error is central to a discussion of dispersion, it should be noted that no mention of the classical theorist's "true score" (Lord & Novick, 1968) is necessary. In Gulliksen's (1950, pp. 212-213) discussion of rating student essay exams, he also avoided the use of true scores in developing his thoughts regarding rater error. He argued that any judge's rating is composed of "the correct score that the student should have received on the content of his paper, if it had not been for reader errors" and simply the "error" made by the reader. His conception of error and that presented here are virtually identical.

Importance of dispersion. If the centroids are the same for different groups of agents, the group which displays least dispersion is most preferred for three related statistical and pragmatic reasons. First, smaller dispersion is associated with greater statistical power when multivariate techniques are used to test for differences between

jobs. If jobs A and B are analyzed by that type of rater which exhibits greatest interrater agreement (i.e., least dispersion), the sensitivity of the statistical test used to identify true differences between the jobs is greatest. The explanation for this sensitivity follows from a consideration of the difference between the jobs on a single dimension.

In the univariate case, differences in mean ratings of the two jobs could be assessed by an F ratio from ANOVA. The greater the difference between the two means relative to the within-group variance, the larger is F. If the difference between the means is held constant and the within-group variance is reduced, F will also be larger and less likely to have resulted from chance.

When considering differences between the two jobs on all dimensions simultaneously, Wilk's likelihood ratio criterion (λ) could be used (Tatsuoka, 1971, p. 85). The logic for testing for differences between centroids parallels that in the univariate case. "The greater the disparity among the several group centroids (relative to within-group generalized variance), the smaller the value of λ" (Tatsuoka, 1971, p. 86). Less generalized variance is then seen to be related to a greater chance of detecting statistically significant differences among centroids. As noted previously, dispersion and generalized variance (Σ) are not identical, but the group with the smallest dispersion will also have the least generalized variance. Therefore, using ratings from the type of rater with least dispersion will result in improved statistical power, all other things being equal.

Second, smaller dispersion is associated with smaller confidence regions about the sample centroid in which the population centroid is

expected to be found (Tatsuoka, 1971, p. 75). (In this discussion, the population centroid is akin to the ratings of the archetypal job if all incumbents in a job perceive the content of the job to be identical and there is no rater error.) Smaller dispersion results in greater certainty about the nature of the job and is, in and of itself, very important. In addition, greater precision in locating the job in multidimensional space facilitates identification of job families. The position of a job relative to other jobs is more clearly represented.

Third, as a practical matter, fewer raters may be used to provide a stable profile of the job across the job dimensions if interrater agreement is high (i.e., dispersion is small). Similarly, when only a certain number of judges may be used, data from raters with high interrater agreement are more useful for determining the equivalence of jobs or in categorizing jobs into families because of the smaller dispersion.

III. SUMMARY

Job analysis agents are considered equivalent if and only if the centroids and the dispersions of the multidimensional distributions of their ratings are the same. The importance of equal centroids becomes apparent when considered in relation to uses of job analysis (e.g., forming job families). Dispersion, defined as the average squared distance from the centroid in standardized space, was also found to be a critical consideration when assessing the interchangeability of raters. For example, using job analysis agents with high interrater agreement (i.e., less dispersion) produces greater statistical power when trying to detect differences between jobs.

This chapter has been devoted exclusively to developing and justifying a definition of equivalence of job analysis agents. Little has been said about how to statistically assess the equivalence of rater types in terms of centroids and dispersions. The discussion of the exact statistical procedures designed to test equivalence as defined above is presented in Chapter V.

CHAPTER II

EXISTING METHODS OF ASSESSING SIMILARITY

Consider the ratings from two judges on several dimensions of a stimulus. The average of each judge's ratings across the dimensions could be computed. In this instance, agreement between the judges would simply be assessed by comparing the similarity of the two numbers representing their mean ratings. However, this approach disregards the pattern of each judge's ratings of the stimulus. Average ratings could be identical even if the two judges strongly disagreed in their ratings of the individual dimensions. Figure 2 demonstrates this possibility by assuming that Judge A has a pattern of (1, 5, 1, 5) and Judge B's pattern is (5, 1, 5, 1). The mean rating for both judges is 3.0. Clearly the pattern of ratings must be considered when determining interrater agreement.

Two approaches are used to represent patterns of ratings (or scores) across several dimensions. First, profile analysis or pattern analysis usually considers ratings to be represented as displayed in Figure 2 in which a separate profile is shown for each judge. Alternately, if persons A and B were being evaluated (instead of providing the evaluations), as in taking a test with four subscores, the two patterns would show how well both did on each part of the test.

A multidimensional orientation is adopted in the second method of representing patterns of ratings. For example, the sequence of ratings

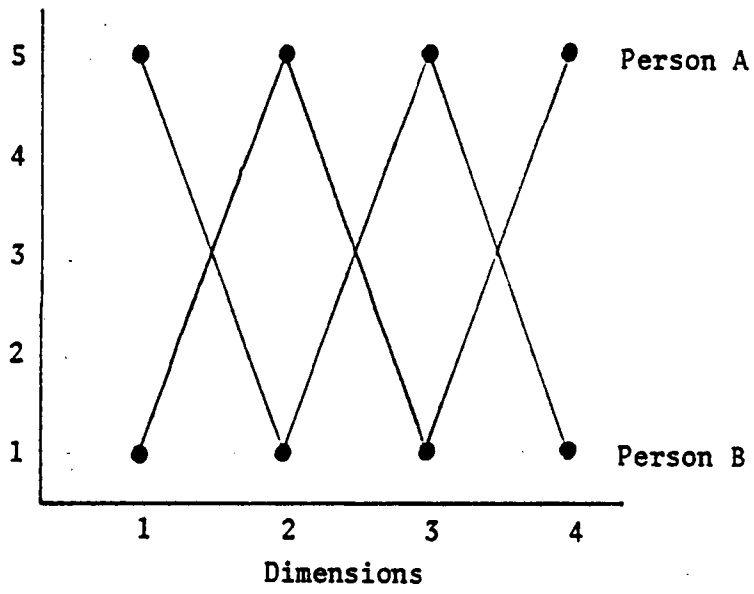


Figure 2. Representing patterns of ratings from two judges.

by Judge A (1, 5, 1, 5) is used as the coordinates of a single point in four dimensional space, each rated dimension serving as a separate axis. Physical representations beyond three dimensions are impossible, but mathematically p coordinates define a single point in p-dimensional space, where p may adopt the value of any positive integer.

No information is lost from the original sets of ratings whether the profile analysis or the multidimensional approach is used. In fact, either manner of representation may be produced from the other. Whichever is used, interrater agreement can be assessed by the similarity of the patterns of ratings. In this chapter, various methods are reviewed for determining similarity. But first, distinctions between levels of comparison need be clarified.

Gregson (1975, pp. 16-20), in his discussion of similarity and levels of judgment, developed a notation for identifying different levels of comparison. L1 denotes the first level of comparison and refers to pairwise comparison judgments. The similarity between two patterns is an L1 comparison; assessing interrater agreement is an L1 comparison.

L2 indicates the second level of comparison and applies to relative pairwise comparisons, as introduced by Gregson. Determining whether two raters agree (i.e., patterns are more similar) more than two different raters agree is an L2 comparison. For example, are the job analysis ratings provided by two job incumbents more similar than the ratings given by two supervisors? For purposes of this exposition, however, L2 is expanded to include comparisons between bwo or more sets of pairwise judgments. Determining whether pairwise agreements among several

among several supervisors or among several incumbents will be considered an L2 comparison. Restated, L2 comparisons are designed to detect differences among similarities.

The review of the relevant literature concerning methods of measuring similarity is categorized according to its level of comparison. The presentation of methods of assessing pairwise similarity (L1) is followed by a review of the few studies which discuss L2 comparisons.

I. REVIEW OF METHODS USED TO DETERMINE PAIRWISE SIMILARITY (L1)

The ability to determine if two configurations of scores are essentially the same is of great importance to psychologists. The two profiles may represent pretreatment and posttreatment scores in which case high similarity indicates an unsuccessful treatment. Similarity in the two patterns may demonstrate high reliability if the profiles are obtained in a test-retest manner. Similarity between a patient's profile and the standard profile of a known mental disorder may facilitate diagnosis. Similarity in the ratings given by two judges indicates high agreement.

Numerous methods have been developed for assessing the similarity between two patterns of scores. As suggested above, the methods were created to address a variety of issues confronting psychologists. However, their usefulness is not limited to the domains of their origins. Methods developed to assess one problem in similarity may be applied to other problems in which similarity is of key importance.

The methods used to develop pairwise similarity may be identified as belonging to one of two categories. First, the correlational approach encompasses many of the methods. The remaining methods are related to some type of distance formulation. That these two categories are not mutually exclusive is not surprising since, under the circumstances, similarity measures based on the correlation between profiles are identical to those which use distance in multidimensional space as their basis.

Correlation as the Basis of Similarity

Pearson r and related methods. The Pearson product-moment correlation coefficient is a long-recognized measure of the degree of association between two variables across people. A high correlation coefficient indicates that the two variables when standardized covary about their respective means.

However, sometimes a psychologist is interested in measuring the degree of similarity between persons and not between variables. For this purpose, Q-type correlation has been advocated by several authors, the most fervent advocate being Stephenson (1935, 1936a, 1936b, 1952). Q-type correlation refers to correlation between persons across variables.

As Cronbach (1953, p. 381) pointed out, "Correlating one person with another isn't half as exotic as it sounds. We do it all the time." He continued by demonstrating that a test score is simply a measure of the agreement between a professor's key and the student's answers. "This scoring operation is logically identical to correlating the two answer sheets, even though we do not express the score as a correlation."

Once one makes the mental transition from thinking of correlating variables to correlating persons, the statistical manipulations involved pose no new difficulties. Transposing the traditional data matrix with persons as rows and variables (e.g., items) as columns before analyzing the data will result in Q-type analyses.

Within the context of this paper, the most obvious application of correlating persons is to measure interrater agreement—assessing to what extent the ratings of two judges covary across rated dimensions. Examining "interrater reliability" in this manner is a common practice among psychologists.

Howard and Diesendaus (1967) called the Pearson r "the most frequently used method for estimating the similarity between two profiles" (p. 225). Nevertheless, correlation between persons as a measure of similarity has been strongly criticized. Gregson (1975) called it a "statistically illiterate practice" (p. 52). The following criticisms are directed at the Pearson r. However, the weaknesses are shared by other correlational techniques which can be used to assess profile similarity, viz., Spearman's rho, Kendall's tau, and du Mas's (1946, 1947, 1949, 1950) coefficient of profile similarity which approximates tau (Cronbach & Gleser, 1953).

Burt (1941, pp. 173-182) summarized criticisms raised by Thomson concerning the use of r (Thomson, 1935, 1939; Thomson & Bailes, 1926). Thomson's primary objection involved the possible disparity of units of measurement when r is computed between persons. For example, suppose that a number of physical characteristics (e.g., arm length, weight, height) were recorded for two individuals and the physical similarity

between them was assessed by r. A data matrix might be prepared in which each row represented a physical characteristic and each column represented a person. In correlating the two persons, the initial step is to calculate the mean score for each person so that the scores within a column will be standardized. However, this process involves averaging scores from a number of different sources. Is it meaningful to calculate means across dimensions for an individual if the units of measurement are dissimilar (e.g., inches, pounds, feet)? Burt (1941) answered this criticism by proposing that the entire problem be avoided by first standardizing the observations within each dimension (i.e., row) before r is calculated which involves standardizing scores within columns.

More serious problems have been raised. A variety of authors has noted that a correlational technique examines differences in only the shape of the profiles and not their levels (e.g., Cattell, 1949; Cronbach, 1958; Cronbach & Gleser, 1953). In discussing this topic, the terminology used by Cronbach and Gleser (1953) to describe the three aspects of a profile is helpful.

> Elevation is the mean of all scores for a given person. Scatter is the square root of the sum of squares of the individual's deviation scores about his own mean. . . . Shape is the residual information in the score set after equating profiles for both elevation and scatter [p. 460].

Figure 3 illustrates these concepts by showing examples of profiles with equal shapes which differ in elevation, scatter, or both. The correlation between persons begins by standardizing scores within persons, setting the elevation in each profile to 0.0 and the scatter to 1.0. It is hardly surprising that correlation can only compare the shapes of the profiles since the method equates profiles on the remaining aspects.

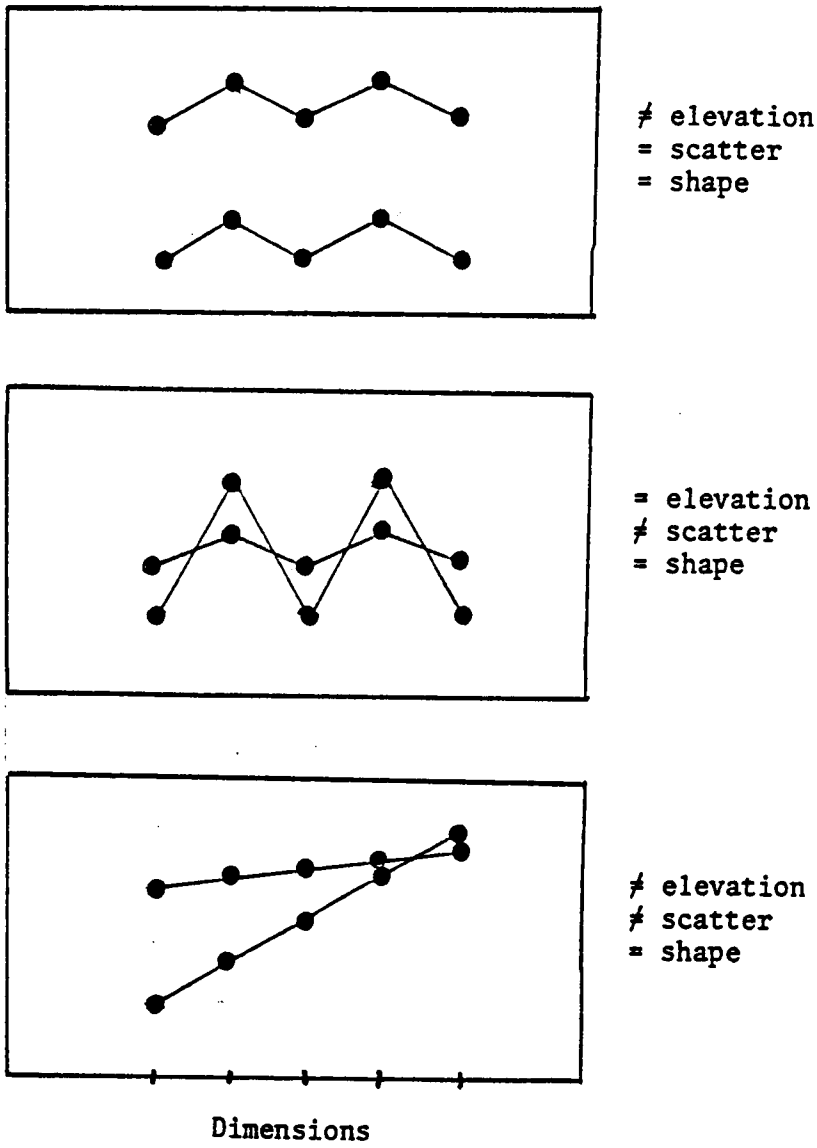


Figure 3. Profiles with equal shapes ($\underline{r}$ = 1.0).

From Figure 3 it can be seen that equal shapes (i.e., r = 1.0) do not imply that the profiles are parallel. If the profiles represent ratings from different judges, one inadequacy of the correlational approach becomes apparent. Judgments which are intuitively quite different are evaluated as being in perfect correspondence. Since all are equally correlated, this approach views two coincident profiles as being no more similar than any of the pairs illustrated.

An extreme example of how the use of correlation may produce nonsensical results is shown in Figure 4. Judges A and B (r = -1.00) are determined to be in extreme disagreement even though their ratings differ very slightly. Using this method, no conceivable pair of profiles can exhibit greater dissimilarity.

Judges C and D (r = 1.00) demonstrate perfect interrater agreement even though their ratings are quite different. In some circumstances, a psychologist might be interested only in comparing the shapes of profiles (see, e.g., Cattel, 1949), but shape alone is clearly inadequate as a criterion for assessing interrater agreement.

Another problem with using correlations between persons, first noted by Cronbach and Gleser (1953) and elaborated on by Tellegen (1965), concerns the direction of measurement. Tellegen (1965, p. 233) begins his discussion of the problem by describing direction of measurement:

> Psychological tests can be said to have a direction of measurement. A scale's direction of measurement determines how high and low scores are to be interpreted. High scores on a certain scale may, for example, be indicative of extraversion. One can, however, reverse its direction of measurement, or, using the more common term, reflect the scale. Now, if the scale is bipolar, high scores are indicative of introversion. Following convention the scale would now be called a measure of introversion.

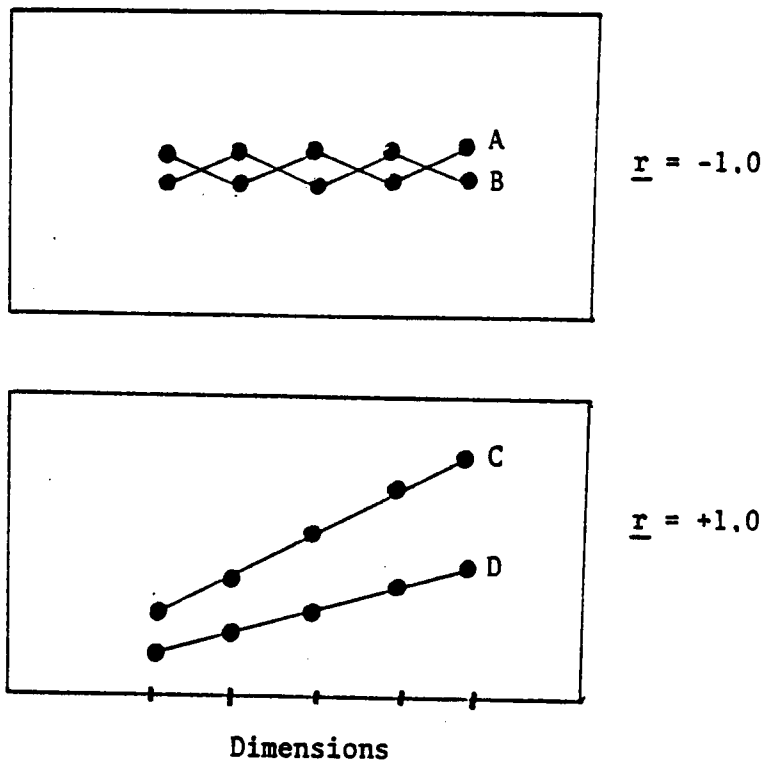


Figure 4. Potentially misleading results from correlational approach.

> Reflection of a scale has no effect on its function of making discriminations on certain dimensions. One might use the following terminology: reflection alters a scale as a directional measure, but leaves it invariant as a differential measure, that is, as a tool for differentiating responses and individuals. If one views a scale as a differential tool, then its direction of measurement is arbitrary, and merely a matter of convention or personal prefer e nce.

Following Tellegen's lead, Howard and Diesendaus (1967) demonstrated how scale reflection can affect the configuration of a profile by altering its elevation, scatter, and shape. (In 1953, Cronbach and Gleser had anticipated this problem.) Most important to the present discussion is the possibility of change in shape, to which correlation is sensitive. Howard and Diesendaus (1967) provided the following remarkable demonstration:

> Consider the following two profiles consisting of sten scores . . . for four traits:
>
> | Trait | Person A | Person B |
> |---|---|---|
> | Outgoing | 9 | 6 |
> | Tense | 8 | 7 |
> | Assertive | 7 | 8 |
> | Venturesome | 6 | 9 |
>
> The correlation between these two profiles is -1.00. How, let us reverse the direction of measurement of the first trait thus (reverse score equals 11 minus original score—still in sten units):
>
> | Trait | Person A | Person B |
> |---|---|---|
> | Reserved | 2 | 5 |
> | Tense | 8 | 7 |
> | Assertive | 7 | 8 |
> | Venturesome | 6 | 9 |
>
> The product-moment correlation between the two profiles is now +.69! [p. 227]

Cronbach and Gleser (1953, p. 464) indicated that unless the dimensions of the profiles are systematically correlated there is generally "no particular reason for scoring each variate in one direction rather than the other." When systematic correlation does exist, all

variables should be scored such that the intercorrelations among variables are all positive (see, e.g., Haggard, 1958). However, this advice offers little help to a researcher comparing profiles, the dimensions of which are based on some orthogonal factor analytic method or when the correlations among the variables fluctuate from sample to sample.

Consider the Managerial Position Description Questionnaire (MPDQ; Tornow & Pinto, 1976), which is composed of 13 orthogonal factors. Since the intent of the MPDQ is to differentiate among jobs, Tellegen would argue that the direction of measurement of the 13 factors is arbitrary. A researcher examining the similarity of the ratings (i.e., factor scores) on the 13 dimensions made by two persons in the same job would have a problem if a correlational method were used. The size (and even the sign) of r might be determined by the researcher's choice of the direction of measurement which is, after all, "merely a matter of convention or personal preference" (Tellegen, 1967, p. 227).

Sjöberg and Holley (1967) and Cohen (1969) have developed correlation-based coefficients which are invariant over variable reflection. Their contributions were in response to the direction of measurement deficiency inherent in the Pearson r. Thus, there does appear to be a solution to this particular correlational problem, although the solution involves using complex coefficients instead of r.

In passing, it should be noted that this particular deficiency is not always present when the direction of measurement is arbitrary. Emphasizing a point made by Cohen (1969), Block (1970) stated that "when the means of the two profiles being related are each at the overall

profile mean" (p. 307) the Pearson r is invariant over variable reflection. That is, under certain restrictive circumstances, the arbitrary choice of direction of measurement creates no problem for correlations.

A final problem which must be raised concerns the interpretation of the magnitude of a similarity index (Cronbach & Gleser, 1953). The magnitude of an index has no meaning in and of itself. Knowing that an individual received a 75 on an exam does not allow meaningful interpretation of the score. Information concerning the difficulty of the items or how 75 compares to other test scores is needed. Correlation between persons has the same problem in interpretation. A correlation equal to .65 between two people is uninterpretable. The size of the correlation is dependent on the difficulty of the items.

Cronbach and Gleser (1953) demonstrated the inadequacy of interpreting correlation coefficients as measures of similarity without regard to the mean value of the dimensions. They re-examined a published study in which the author showed that the Rorschach test cannot be faked. According to Cronbach and Gleser, the author of the study had:

> . . . hoped to demonstrate that the Rorschach test is proof against faking. He therefore asked individuals to take the test in the normal manner, and then to take it attempting to make the best possible impression. He correlated the two psychograms for a given individual and interpreted the resulting high correlations as showing that the Rorschach was proof against faking. Now it is true that the psychogram under "fake good" conditions could be predicted from that under normal conditions. But the "fake good" psychogram could have been predicted quite well from the psychogram of some other person chosen at random. Between any two Rorschach records taken at random, there will tend to be a high correlation just because certain scores (e.g., D, F) will usually be large, and other scores (e.g., m, cF) will usually be small [p. 458].

Although not discussed by Cronbach and Gleser (1953), the problem of finding a high interrater correlation because of widely divergent dimension means can be avoided by standardizing the variables. However, still remaining is their more basic point—the interpretation of correlation coefficients (or any measure of profile similarity) should be carefully considered in the context in which it was calculated.

Intraclass correlation. In 1951, Ebel discussed the uses of the intraclass correlation techniques for estimating the reliability of ratings. More recently, Shrout and Fleiss (1979) described a number of different types of intraclass correlations which can be used to assess rater reliability. Worth mentioning is that the intraclass correlation is "a special case of the one-facet generalizability study (G study) discussed by Cronbach, Gleser, Nanda, and Rajaratnam (1972)" (Shrout & Fleiss, 1979, p. 420).

Any intraclass correlation coefficient is a ratio of the variance of interest over the sum of the variance of interest plus error. The relationship to reliability becomes apparent when reliability is defined as the ratio of true score variance to the sum of true score variance and error variance (Lord & Novick, 1968). Both the intraclass correlation and reliability are determined by variance ratios.

In a typical design with $\underline{r}$ raters judging $\underline{d}$ dimensions, the variance components due to raters ($\hat{\sigma}^2_{\underline{R}}$), dimensions ($\hat{\sigma}^2_{\underline{D}}$), and residual error ($\hat{\sigma}^2_{\underline{e}}$) may be identified. Assuming no interaction, the obtained MS_{rater} in an ANOVA summary table estimates $\hat{\sigma}^2_{\underline{e}} + \underline{d}\ \hat{\sigma}^2_{\underline{R}}$. $MS_{dimension}$ estimates $\hat{\sigma}^2_{\underline{e}} + \underline{r}\ \hat{\sigma}^2_{\underline{D}}$.

MS_{error} estimates $\hat{\sigma}^2_e$. To calculate $\hat{\sigma}^2_e$, $\hat{\sigma}^2_R$, and $\hat{\sigma}^2_d$ involves simple arithmetic:

$$\hat{\sigma}^2_e = MS_{error}$$

$$\hat{\sigma}^2_R = (MS_{rater} - MS_{error})/d$$

$$\hat{\sigma}^2_d = (MS_{dimension} - MS_{error})/r$$

To determine the reliability of ratings, one need only compute

$$r_{intra} = \frac{\hat{\sigma}^2_d}{\hat{\sigma}^2_d + \hat{\sigma}^2_e} \qquad (1)$$

Formula 1 will adequately assess the similarity of the shapes of the r profiles; differences in level are ignored. This method of determining similarity shares this problem with the Pearson r as discussed earlier.

Unlike the Pearson r, however, the intraclass correlation coefficient may be modified to treat differences in rater level as error:

$$r_{intra} = \frac{\hat{\sigma}^2_d}{\hat{\sigma}^2_d + \hat{\sigma}^2_e + \hat{\sigma}^2_R} \qquad (2)$$

Ebel (1951) described the issue concerning whether differences in level should be allowed to affect the intraclass correlation as follows:

> Whether or not it is desirable to remove "between-raters" variance in estimating the reliability of ratings depends upon the way in which the ratings are ultimately used in grading, classification, or selection. In any case, where differences from rater to rater in general level of rating do not lead to corresponding differences in the ultimate grades, classifications, or selections, the "between-raters" variance should be removed from the error term. . . . But

> if decisions are made in practice by comparing single "raw" scores assigned to different pupils by different raters, or by comparing averages which come from different groups of raters, then the "between-raters" variance should be included as part of the error terms [pp. 411-412]

It should be clear from the earlier discussion and examples presented in Figures 2 (p. 15) and 3 (p. 21) that differences in rater level indicate dissimilarity within the context of job analysis ratings. Therefore, between-rater variance is best treated as a source of error in computations of the intraclass correlation used to assess profile similarity.

Although the intraclass correlation approach deals satisfactorily with the "level" problems, this method is still susceptible to the "direction of measurement" problem. The often arbitrary decision made by the researcher as to the direction a variable will be scored affects the value of the intraclass correlation coefficient. Analysis of the data set shown in Table 1 illustrates this point. Formulas 1 and 2 were used to determine the reliability of the ratings in their original form. However, nonsensical results ($\hat{\sigma}^2_R < 0$) were found in the data set after one of the scales was reflected. As can be seen, the intraclass correlation is affected by scale reflection.

Despite the weaknesses noted here, the intraclass correlation is superior to the Pearson r for assessing profile similarity because of its sensitivity for detecting differences in level or elevation.

Distance Formulations

As was described in the chapter introduction, the pattern of ratings from a judge may be represented as a two-dimensional profile of scores such as shown in Figures 2, 3, and 4 (pp. 15, 21, and 23), or as a sinble point in multidimensional space. Either manner of

TABLE 1

DIRECTION OF MEASUREMENT AFFECTS $\underline{r}_{INTRA}$

Data Set	Dimension	Rater A	Rater B	ANOVA Summary Table Source	df	SS	MS
Original Data[a]	a	3	6	Rater	1	12.5	12.5
	b	2	5	Dimension	3	10.5	3.5
	c	5	7	Error	3	.5	.1667
	d	2	4	Total	7	23.5	
After Scale Reflection[b]	a	5	2	Rater	1	2	2.0
	b	2	5	Dimension	3	11	3.6667
	c	5	7	Error	3	11	3.6667
	d	2	4	Total	7	24	

[a] $\underline{r}_{intra}$ = .91 (Formula 1); $\underline{r}_{intra}$ = .34 (Formula 2).

[b] $\underline{r}_{intra}$ = 0.0 (Formula 1); since $\hat{\sigma}_{\underline{R}}^2 < 0$, $\underline{r}_{intra}$ is not calculated (Formula 2).

Dimension a is reflected (7-point scale).

conceptualization may be created from the other. Nevertheless, these two approaches have led to very divergent solutions to the problem of how to assess the similarity between two patterns of ratings. In general, developers of correlational approaches thought in terms of the two-dimensional array. Once the mental set of researchers shifted to thinking in terms of multidimensional space, they naturally developed similarity measures based on distance. Their reasoning was that the more similar any two patterns of ratings the closer they will be in the multidimensional space.

Cronbach and Gleser (1953) carefully explored and developed the concept of assessing profile similarity via distance measures. They introduced the following notation (p. 459):

j = any of the variates a, b, c, . . . which are k in number;
i = any one of the persons 1, 2, . . . N;
x_{ji} = the score of person i on variable j.

In the discussion to follow, x_{ji} may also equal the scores <u>given</u> by the i^{th} person on variate j. The method of measuring distance which is explained below, is inherent in the Cronbach and Gleser work, even though they avoided the use of matrix algebra.

Using matrix notation, the set of ratings given by person 1 may be represented as a column vector with <u>k</u> elements:

$$\underline{X}_1 = \begin{bmatrix} x_{a1} \\ x_{b1} \\ x_{c1} \\ \vdots \\ x_{k1} \end{bmatrix}$$

The end of this vector identifies a single point. In a like fashion, the end of $\underline{X}_2$ designates a point associated with a second judge's ratings. Similarity between the two sets of ratings is assessed by computing the linear distance between the respective points. If the variables are represented by orthogonal axes, the squared distance between the points is equal to:

$$\underline{D}^2 = (\underline{X}_1 - \underline{X}_2)'(\underline{X}_1 - \underline{X}_2) \qquad (3)$$

This expression is equivalent to the sum of the squared differences on each variable. For a more detailed presentation of how to use the Pythagorean theorem to determine distances in $\underline{k}$ dimensions, see Coombs, Dawes, and Tversky (1970, pp. 371-377).

Either $\underline{D}^2$ or $\underline{D}$ may be used as a measure of similarity. Cronbach and Gleser (1953, p. 459) preferred $\underline{D}$ because "larger differences between persons are much exaggerated in squaring" and "$\underline{D}$ is less skewed than $\underline{D}^2$." However, they also noted that normality is not achieved by using $\underline{D}$, only less skewness.

The cause of this skewness is readily seen if one imagines a multivariate normal distribution of points in $\underline{k}$-space. Most of the points will be fairly close together, but some pairs will be quite distant. The smallest possible value of $\underline{D}$ is zero, but $\underline{D}$ can be infinitely large. For variables of this type, a positive skew is common (Chou, 1975, p. 72). If $\underline{D}^2$ were computed, the skewness of the distribution would be even more accentuated. If the degree of skewness is not a consideration, the choice between $\underline{D}$ and $\underline{D}^2$ seems to be largely a matter of personal preference.

$\underline{D}$ (or $\underline{D}^2$), as a measure of similarity, avoids many of the problems associated with Pearson $\underline{r}$ or the intraclass correlation. $\underline{D}$ is invariant over scale reflection and simultaneously assesses differences in elevation and shape. A simple example is one dimension will illustrate the former.

Suppose two movie patrons decide to rate the attractiveness of the film's leading lady, using a 10-point scale with 10 being "most attractive." The first rater views her as a "10" whereas the second judge, considerably less impressed, assigns her a score of "7." The squared difference between the two ratings is 9 (i.e., $(10-7)^2 = 9$). If the attractiveness scale were reflected such that most attractive is given a value of "1," the rating from the first judge would be "1" and a "4" would be given by the second judge. The squared difference between the ratings remains 9. Reflection of scale values does not alter the squared differences between two values. Since $\underline{D}^2$ in $\underline{k}$-space is equal to the squared differences between the values on each dimension summed over the $\underline{k}$ dimensions, reflection of any or all variables does not affect its value.

$\underline{D}^2$ is sensitive to differences in all characteristics of a profile (viz., elevation, scatter, and shape) because all the available pattern information is used to construct the vectors between which distance is measured. This is certainly a refinement over $\underline{r}$. Yet, Cronbach (1958) discussed methods for isolating different components of profiles (e.g., elevation) which may then be compared individually. He described situations in which these finer techniques might be appropriately applied. In the present context of examining the similarity of raters'

judgments, any type of difference in the profiles represents disagreement. Because of this, $\underline{D}^2$ is preferred as an overall measure of the correspondence between two sets of ratings to the other measures outlined by Cronbach (1958).

Cronbach and Gleser's (1953) practice of representing the variables by orthogonal axes has been challenged. They stated that $\underline{D}^2$ may be computed "whether variates are correlated or not" (p. 470). Overall (1964) took strong exception to their remark:

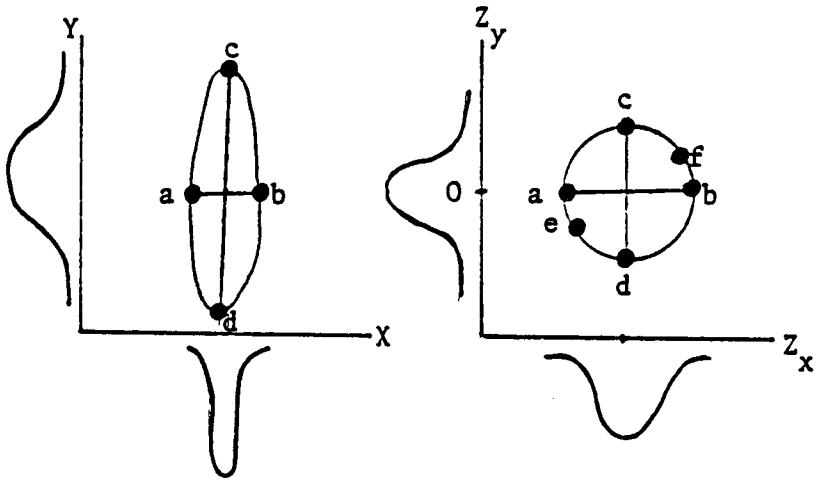
> Distance is defined as the square root of the sum of the squared differences in projections on uncorrelated reference axes. If correlations between profile elements exist, then the distance between two profiles in multidimensional space is not the square root of the sum of squared differences in scores on the correlated profile components. If differential or heterogeneous intercorrelations exist, a very distorted picture of true distances between individuals may result from the simple summing of squared differences on the separate measures [p. 196].

Overall advocated use of the Mahalanobis generalized distance measure. The formula for computing the distance between two points is:

$$\underline{d}^2 = (\underline{X}_1 - \underline{X}_2)'\hat{\Sigma}^{-1}(\underline{X}_1 - \underline{X}_2) \qquad (4)$$

where $\underline{d}^2$ is used to distinguish the Mahalanobis distance from the Pythagorean $\underline{D}^2$, and $\hat{\Sigma}^{-1}$ is the inverse of the variance-covariance matrix for the $\underline{k}$ profile elements.

The purpose of the Mahalanobis concept is to produce a standardized space such that distance in every direction is measured in terms of some standard deviation. Consider the scatter plots shown in Figure 5. Variables x and y are uncorrelated and have unequal standard deviations. The Pythagorean distance between a and b (denoted $\underline{D}_{ab}$) is less than $\underline{D}_{cd}$.



$r_{XY} = 0$
$\sigma_X^2 \neq \sigma_Y^2$

$r_{XY} = 0$
$\sigma_X^2 = \sigma_Y^2$

Figure 5. Effects of standardizing space.

However, a and b are separated by about four standard deviations on x, and c and d are separated by about four standard deviations on y. The meaninfgulness of the differences is the same, although the values of $\underline{D}$ are quite different. Standardizing the variables x and y before computing $\underline{D}$ changes the results dramatically. $\underline{D}_{ab}$ now equals $\underline{D}_{cd}$, which is congruent with their equivalent meanings.

Since x and y were uncorrelated initially, standardizing the variables produced a standardized space in the Mahalanobis sense. The distance between any two points (e.g., $\underline{D}_{ef}$) is being expressed in units comparable to any other distance. When the $\underline{k}$ variables in $\underline{k}$-dimensional are correlated, the Mahalanobis formula still works; distances are expressed in equivalent terms.

By returning to Formula 4, it can be shown that when variables are uncorrelated and standardized, $\underline{D}^2 = \underline{d}^2$. For this case, Overall (1964) observed that $\hat{\Sigma}^{-1}$ will be proportional to an identity matrix in which case Formula 4 reduces to the Pythagorean distance formula, $\underline{D}^2 = (\underline{X}_1 - \underline{X}_2)'(\underline{X}_1 - \underline{X}_2)$. The entire problem is avoided if the dimensions are standardized and uncorrelated as might be found in dimension ratings based on orthogonal factor scores.

Heerman (1965) argued that the orthogonality or obliqueness of coordinate axes is an arbitrary choice made by the researcher and, therefore, the Pythagorean theorem could be used even with correlated profile elements. Nunnally (1967) concurred with this conclusion when he stated

> There is no mathematical necessity for restricting the use of $\underline{D}$ . . . to those situations where variables are

> uncorrelated. For the mathematical analyses, an orthogonal space is constructed, and there is no need to make the angles among vectors proportional to their natural correlations. The real issue concerns the interpretability of results [pp. 385-386].

Because of the issue of interpretability, the Mahalanobis d^2 approach is more desirable than the Pythagorean theorem distance formula. However, when d^2 cannot be computed because of inadequate information concerning the within-group variance-covariance matrices, the arguments presented by Heerman (1965) and Nunnally (1967) support the use of D or D^2.

Cronbach and Gleser (1953) compared D^2 with other measures of profile similarity. Two earlier distance-based measures, Pearson's (1928) "coefficient of racial likeness" and Osgood and Suci's (1952) approach, were recognized. Cattell's (1949) "coefficient of pattern similarity" (r_p) was discussed within the context of their own distance formulations. Their primary objection to r_p was that the values for the coefficient are restricted to the range of +1.0 to -1.0, where -1.0 represents maximum dissimilarity. Cronbach and Gleser (1953) argued that complete dissimilarity of persons is an undefinable concept. For any pair of points separated by a given distance, there can always be defined another pair more dissimilar. The restriction of any coefficient to a finite interval reduces the utility of the coefficient as a measure of similarity. Nonetheless, a table of the standard error for r_p (Horn, 1961) has increased the popularity of the statistic.

Cronbach and Gleser (1953) continued their review by demonstrating that Webster's (1952) version of the intraclass correlation is inadequate for assessing similarity. The formula for Webster's intraclass correlation is

$$\underline{r}_{in} = 1 - \frac{\underline{D}_{12}^2}{\underline{S}_1^2 + S_2^2 + \frac{1}{2} K\Delta^2 El}$$

where $\underline{D}_{12}$ = the Pythagorean distance between persons in k-dimensional space, $\underline{S}$ refers to the scatter of an individual's scores (i.e., the standard deviation within the profile), and ΔEl refers to difference in elevations between the profiles. Cronbach and Gleser (1953) described the distinction between $\underline{D}$ and $\underline{r}_{intra}$ most clearly. They noted that the denominator in the formula:

> is the sum of squares of scores of both persons about the grand mean of their scores. The larger this denominator the closer will $\underline{r}_{in}$ approach +1 for pairs having the same $\underline{D}$. To illustrate, consider person X with standard scores 1.0, 1.0, on two variates and person with Y standard scores 1.1, 1.1. For this pair, $\underline{D}$ is $\sqrt{.02}$. $\underline{S}$ for each person is zero, the denominator is small, and $\underline{r}_{in} = -1$. In other words, this pair of persons is reported by intraclass $\underline{r}$ to have maximum dissimilarity, whereas the $\underline{D}$ measure reports them to be close together. The definition upon which the $\underline{d}$ measure is based appears to present a more satisfactory conception of similarity than the definition embodied in the intraclass measure [pp. 462-463].

Finally, if scores are standardized within profiles, $\underline{D}$ is equivalent to the Pearson $\underline{r}$. By definition, all that remains after removing elevation and scatter (the product of standardizing within profiles) is shape. Because no other differences remain, $\underline{D}$ can only examine the similarity of the profiles' shapes—a function identical to that performed by $\underline{r}$. Under these circumstances, $\underline{r}$ can be viewed as a special case of $\underline{D}$.

Summary

Methods related to Pearson $\underline{r}$, intraclass $\underline{r}$, and distance for assessing profile similarity were reviewed. Pearson $\underline{r}$ was shown to be

clearly deficient because of its inability to detect differences in elevation, its dependence on direction of measurement, and its finite range of possible values. Intraclass r can treat differences in elevation as error, but r_{intra} shares the other problems listed for Pearson r. Distance measures have none of these deficiencies and appear most promising as means of determining similarity.

II. REVIEW OF METHODS USED TO ASSESS DIFFERENCES AMONG PAIRWISE SIMILARITIES (L2)

The development of a statistical method for assessing differences between groups of raters in terms of their within-group interrater agreement is a primary purpose of this dissertation. As already noted, this type of comparison is of the second level (L2)—the comparison of similarities.

One approach is for the researcher to examine the indices of similarity calculated from the data. Based on the way they "look," the researcher makes implicit or explicit inferences regarding the differences in interrater agreement (e.g., Greenwood & McNamara, 1967; Hazel, Madden, & Christal, 1967; Probst, 1932).

Approaches to comparing profile similarity coefficients have been developed for measures based on Pearson r and the intraclass correlation. Guilford and Fruchter (1978) described a procedure to assess the difference between two independent correlation coefficients by computing Z, a test statistic distributed as a standard score. The probability of occurrence for the calculated value of Z under the null hypothesis of no difference in the coefficients may be readily obtained from a table of

the normal standard distributions. Where the correlation between two profiles has been transformed to a Fisher Z, the difference between $\underline{r}_1$ and $\underline{r}_2$ may be computed as follows:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - N_3} + \frac{1}{N_2 - 3}}} \qquad (5)$$

In the case of correlating persons, $\underline{N}$ refers to the numbers of dimensions being rated. In the notation used earlier, the denominator of the expression is equal to $\sqrt{\frac{1}{K_1 - 3} + \frac{1}{K_2 - 3}}$.

If several values of $\underline{r}$ were calculated within a group which were to be compared to several values of $\underline{r}$ from within another group, the average $\underline{r}$ within each group (based on Fisher transformations) could be entered into Formula 5 (Edwards, 1934). The power of this statistical test is inversely related to the number of dimensions over which persons are correlated. When $\underline{k}$ is small, a nonparametric test of the difference in median $\underline{r}$ would prove more profitable.

Procedures for assessing the difference between two intraclass correlation coefficients have been presented (Haggard, 1958) or may be developed from discussions of $\underline{r}_{intra}$ confidence intervals (Ebel, 1951; Shrout & Fleiss, 1979). The formulas are lengthy and will not be reproduced here. However, it should be noted here that these procedures have little power when the numbers of raters or the number of dimensions is small. Published studies in which this technique is used to determine differences in interrater agreement are rare, if not nonexistent.

No procedure has previously been developed which compares similarity indices defined in terms of distance, although the usefulness of such a procedure is inferred by Cronbach and Gleser (1953). This oversight is surprising, considering the advantages distance has as a measure of profile similarity.

CHAPTER III

REVIEW OF RELEVANT RESEARCH IN WHICH INTERRATER AGREEMENTS ARE ASSESSED

The purpose of this chapter is to review the literature regarding the equivalence of different job analysis agents. However, included in this review are several studies which have nothing to do with job analysis but are germane because they describe approaches by which groups of raters have been compared. Studies reviewed in the chapter are organized according to the statistical methods employed by the researchers.

I. CORRELATION TO ASSESS INTERRATER AGREEMENT

r Between Different Types of Raters

Despite the problems inherent in correlational approaches, the Pearson product-moment r is frequently used to assess interrater agreement. Several recent empirical studies report correlations between different types of job analysis agents rating jobs on the Position Analysis Questionnaire (PAQ). McCormick, Mecham and Jeanneret (1977) computed all possible pairwise correlations among the raters in their sample. They reported that the average correlation ($\bar{r}$) between job analysts and supervisors was .83; between analysts and incumbents, $\bar{r}$ = .84; and between supervisors and incumbents, $\bar{r}$ = .79. Taylor (1978) found only "modest levels of interrater agreement" (p. 336) among

the incumbents, supervisors, and others "familiar with the job" who completed the PAQ. The median r was .46 across the 27 component dimensions and .50 across the five overall dimensions of the PAQ. Taylor and Colbert (1978) reported that the average r for 1,190 pairs of raters was .68. The median r was .59 between the job analysts and the incumbents in the study conducted by Cornelius, Carron and Collins (1979).

Smith and Hakel (1979) used a slightly different approach in computing correlations between different types of raters. Other authors first calculated correlations between all pairs of raters and then reported averages of the correlations. Smith and Hakel (1979) started by calculating the mean of the ratings given to each PAQ dimension by each type of job analysis agent. Pairwise correlations between the types of agents were computed over the means ascribed to the PAQ dimensions. The correlation of mean ratings from different sources (viz., job analysts, incumbents, and supervisors) ranged from .92 to .98. In the Smith and Hakel (1979) study, students given only the job title and other students provided with job specifications also completed the PAQ. None of the correlations of the mean ratings given by these two groups of students and the mean ratings from the three traditional types of job analysis agents was less than .89.

Hackman and Oldham (1974) were interested in the correspondence of ratings from different types of raters when they completed the Job Diagnostic Survey (JDS). In their approach, "The ratings of each group (i.e., employees, supervisors, observers) were averaged for each job, and the correlations were computed using jobs as observations" (p. 19).

In this manner, separate correlations between types of raters were calculated for each job dimension. The median r between employees and supervisors was .51; between employees and observers, the median r was .63; and between supervisors and observers, the median r was .46.

One other study in which correlation was used to assess agreement between different types of raters is that conducted by Lawshe and Farbo (1949). They found that when union members and managers rated the job on such factors as working conditions and job hazards, the pairwise correlations between union members and managers ranged from .66 to .86.

r Among Raters of Same Type

In order to assess the reliability of ratings, correlations between individuals in the same job analysis agent category have been computed. When rating jobs with the PAQ, Smith and Hakel (1979) reported that the average r among incumbents was .59; among supervisors, $\bar{r}$ = .63; among analysts, $\bar{r}$ = .63; among students given only job titles, $\bar{r}$ = .51, and among students given job specifications, $\bar{r}$ = .49. Cornelius and Lyness (1980) used interrater agreement among incumbents, measured by means of r, as one indicator of the quality of their job analysis ratings. Tornow and Pinto (1976) cited Rushmore's (1961) observation that "it is not surprising to find relatively low intercorrelations between position descriptions by supervisors" (p. 412). Finally, Greenwood and McNamara (1967) calculated the reliability of judgments provided by professional and nonprofessional evaluators in assessment centers. Their approach was to compute the average correlation among the professional evaluators and the $\bar{r}$ among the nonprofessional evaluators.

Susceptibility of r to Range Restrictions

The problems associated with using r to assess interrater agreement were presented at length in the preceding chapter. It was demonstrated that the Pearson r is incapable of detecting differences in elevation, dependent on the direction of measurement, and limited to a finite range of possible values. Further, the variability of the mean values of the dimensions over which the correlation is calculated affect r. Cronbach and Gleser (1953) demonstrated this phenomenon in their re-examination of the fakeability of the Rorschach test which was discussed earlier. A high correlation tends to be produced when some mean values are low and some are high. The two pairs of profiles in Figure 6 have equal elevations and are equally distant on each dimension, but the variation of dimension means is greater in the lower pair. The consequence of the increased variation in means is a sizable positive correlation across the dimensions. Smith and Hakel (1979) recognized this problem when they stated, "These high correlations may merely be indicative of the frequency with which the PAQ elements actually occur in the world" (p. 685). In discussing reliability as a correlation of ratings over jobs, Ash (1948) acknowledged that "The great variability of jobs in this sample . . . probably tended to increase the reliability coefficients" (p. 320).

A re-examination of the data presented by Hackman and Oldham (1974) empirically demonstrates the susceptibility of r to change in the mean values of the dimensions. Their findings concerning the relationships among employees', supervisors', and observers' job ratings are reproduced in Table 2. The correlation between employees and supervisors on skill variety (r = .64) was calculated as follows:

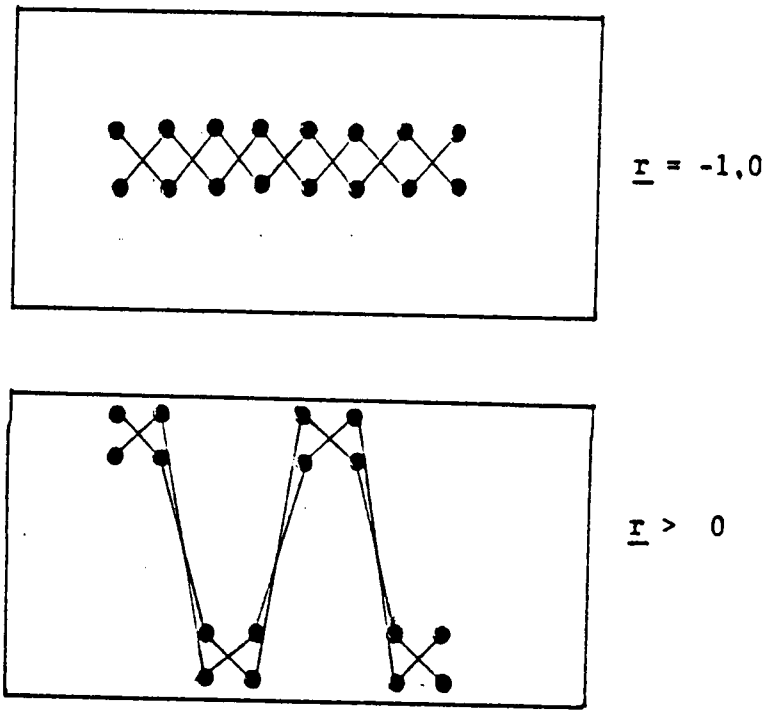


Figure 6. Differences in r produced by differences in the variability of the mean values of the dimensions.

TABLE 2

RELATIONSHIPS AMONG EMPLOYEES', SUPERVISORS' AND OBSERVERS' JOB RATINGS

Job Dimension[a]	Correlations Between		
	Employees and Supervisors	Employees and Observers	Supervisors and Observers
Skill Variety	.64	.66	.89
Task Identity	.31	.32	.44
Task Significance	.48	.65	-.14
Autonomy	.58	.76	.72
Feedback from the Job Itself	.33	.58	.47
Feedback from Agents	.07	-.13	.14
Dealing with Others	.55	.61	.37
Motivating Potential Score	.56	.70	.71
Median	.51	.63	.46

Note: Data are included only for those jobs for which more than one set of supervisory ratings were available. Ns ranged from 12 to 21 jobs.

[a]This table is reproduced from Hackman and Oldham's (1974, p. 21) technical report.

1. The mean of the ratings on skill variety among employees was calculated for job A.
2. The mean of the ratings of skill variety among supervisors was calculated for job A.
3. These first two steps were repeated for several other jobs producing a data matrix with rows representing jobs and the two columns representing the mean skill variety ratings from employees and supervisors.
4. The correlation between the two columns was computed by the Pearson product-moment correlation formula.

Other correlations reported in Table 2 were calculated in a similar manner.

The size of the correlations is dependent on the value of the mean ratings across the various jobs. If the means among the jobs differ widely, the correlation should be high. Interrater correlations should be highest on those job dimensions on which the jobs in the sample vary most.

Fortunately, Hackman and Oldham (1974) provided $\underline{r}$ values which describe the variability of the mean values of each job dimension across the jobs in the sample. Table 3 shows the average interrater $\underline{r}$ (via Fisher $\underline{z}$ transformations), calculated from the values shown in Table 2, and the $\underline{F}$-ratio from ANOVA across jobs for each job dimension. As is observed, higher values of $\underline{r}$ are associated with larger $\underline{F}$ values. In fact, the correlation between $\underline{r}$ and $\underline{F}$ across the eight job dimensions is .65 ($p < .05$). The conclusion to be drawn is that the size of the interrater correlation is a function of the relative mean values of the

TABLE 3

THE RELATIONSHIP BETWEEN r AND VARIABILITY IN JOBS

Job Dimension	Average Interrater r[a]	ANOVA Across Jobs F
Skill Variety	.75	11.49
Task Identity	.35	3.45
Task Significance	.37	2.08
Autonomy	.69	5.11
Feedback from Job Itself	.47	2.51
Feedback from Agents	.03	2.99
Dealing with Others	.52	4.96
Motivating Potential Score	.66	4.85

[a]Based on values in Table 2 using Fisher Z transformations.

dimensions over which the correlation is calculated, and not merely related to interrater agreement. Restriction in range affects correlation between persons in a manner analogous to the way it affects correlation between items.

Intraclass Correlation to Assess Interrater Agreement

The intraclass correlation coefficient has received considerably less attention than the Pearson r as a measure of interrater agreement. Nevertheless, Arvey and Mossholder (1977) recommended its use to guarantee adequate interrater agreement about the nature of a job before trying to identify jobs which are dissimilar. Borman (1979) used the intraclass correlation coefficient to establish that high interjudge agreement existed among the experts in his study. Madden (1964) calculated intraclass correlations for various methods of constructing rating scales. The effectiveness of the methods was then compared in terms of the obtained values of r_{intra} which were interpreted as reliability coefficients. Although statistical tests of differences between intraclass correlation coefficients are fairly familiar to today's researchers, they were less well known at the time of Madden's (1964) study. It is not surprising that Madden's (1964) conclusion is based on a casual comparison of the intraclass correlation coefficients rather than statistical significance levels.

A study by Harding, Madden, and Colson (1960) elucidated the difference between the intraclass and the product-moment correlation coefficients. While examining job evaluation ratings,

> Two measures of reliability were computed. The correlational approach provided fairly high values, while estimates based on components of variance were lower [p. 357].

They found that whereas r ranged from .57 to .80, the intraclass correlation coefficient calculated from the same data was never higher than .52 and fell as low as .02.

II. COMPARISON OF INTERRATER AGREEMENTS

Researchers have occasionally been interested in determining whether interrater agreement is greater within one group of judges than within another. For example, in the earliest empirical study of job analysts, Probst (1932) found that:

> Of the five judges who are in greatest disagreement with the average opinion, four of them had no experience in hiring or supervising employees; while of the five judges who agreed most closely with the average opinion of the thirty judges, four of them were experienced in supervising employees [pp. 648-649].

In Greenwood and McNamara's (1967) investigation of professional and nonprofessional assessment center evaluators, they stated:

> When one compares in a given situation the heterogeneous group (professional and nonprofessional) reliabilities to the homogeneous group (professional and professional, or nonprofessional) reliabilities, they generally appear equivalent. Also, neither homogeneous group consistently appears to have greater interrater reliability than the other homogeneous or heterogeneous groups [p. 105].

Lawshe and Farbo (1949) concluded:

> In comparing relative reliabilities of management and labor committee members, those of management were found consistently higher (ranging from .80 to .94); than those of labor (ranging from .37 to .83). The magnitude of reliability or agreement of average intercorrelations of the management and labor union

> committee members combined (range from .66 to .86 for one rater) falls between those of labor union members which are of lowest magnitude and those of management representatives which are of highest magnitude [p. 165].

Cornelius and Lyness (1980) were interested in the effects of judgment strategy (holistic, decomposed-clinical, and decomposed-algorithm) on interrater agreement. In describing their findings concerning the superiority of the decomposed-algorithm strategy, they stated:

> In 8 of 10 jobs, the average correlations among incumbent ratings were highest in this treatment. In addition, in 5 of 10 jobs, the decomposed-clinical condition had the lowest average correlation, and in 5 of the jobs, the holistic condition had the lowest correlation [p. 160].

In each of these four studies (Cornelius & Lyness, 1980; Greenwood & McNamara, 1967; Lawshe & Farbo, 1949; Probst, 1932), explicit or implicit inferences about interrater agreement were made. These inferences were based solely on a casual inspection of the data without benefit of rigorous statistical analyses. The only study in which statistical techniques were used to test for differences between groups in terms of within-group interrater agreement was conducted by Smith and Hakel (1979). Mean reliability coefficients (average interrater $\underline{r}$) were computed for each type of rater. The authors stated no significant differences were found between the $\underline{\bar{r}}$'s. Although they do not state exactly how these statistical tests were conducted, probably a test for differences between coefficients or correlation which are uncorrelated (e.g., Guilford & Fruchter, 1978, pp. 163-164) was performed for every pair of correlation coefficients.

III. TESTS OF DIFFERENCES IN MEANS

The definition of equivalence presented earlier requires that the centroids of ratings from different groups of raters not be significantly different. Although no previous research has adopted this multivariate perspective, some literature exists in which researchers have tested for mean differences in the ratings given by different types of raters. The authors have either tested for differences in elevation; i.e., the mean of all item means (Smith & Hakel, 1979; Wiley & Jenkins, 1963), or performed separate independent univariate tests on means (Arvey, Passino, & Lounsbury, 1977; Greenwood & McNamara, 1976; Hackman & Oldham, 1974; Heneman, 1974).

Smith and Hakel (1979) calculated the average value of 24 socially desirable items for each of the five types of raters in their study. An ANOVA F was significant, indicating differences among the groups of judges in the way they responded to these items. No significant differences were found using the mean of the unimpressive PAQ items as the dependent variable. With respect to the PAQ taken in its entirety, Smith and Hakel (1979) concluded that "Supervisors and incumbents seem to rate all or most PAQ items higher than their analyst counterparts. . . . Analysts rated PAQ items significantly lower than all other groups, demonstrating the relative leniency of other groups in using the PAQ response scales" (p. 688). Unfortunately, they did not report the statistical procedures used to reach these conclusions.

Hackman and Oldham (1974, p. 30) used ANOVA to test for mean differences in the ratings given by employees, supervisors, and observers on each of the JDS job dimensions. On all but two of the job dimensions,

on each of the JDS job dimensions. On all but two of the job dimensions, the mean ratings were significantly different among the groups ($p < .05$).

Greenwood and McNamara (1967) compared the mean ratings given by professional and nonprofessional evaluators on several assessment center exercises. Only one of the 15 *t*-tests was significant. "The comparison between professional and nonprofessional evaluators indicated that their ratings are about equivalent for a specific group of participants being assessed" (pp. 105-106).

Heneman (1974) examined differences in self and superior ratings of performance. Significant mean differences were found on four of the nine performance dimensions.

Arvey, Passino, and Lounsbury (1977) investigated the effects of sex of analyst. Of 32 dimensions of the PAQ, significant ($p < .01$) sex effects were found on only one (evaluating information from things).

IV. TESTS OF DIFFERENCES IN VARIANCES

In only one study which was reviewed did the author conduct univariate tests of differences in variances, although such tests would indirectly assess the equality of dispersions. As argued in Chapter II, equality of centroids and dispersions is necessary for equivalence. Heneman (1974) reported significant differences ($p < .05$) in the variances of ratings given by self and superior on four of the nine performance dimensions.

Arvey, Passino, and Lounsbury (1977) found few mean differences on the PAQ dimensions between male and female analysts. However, further analysis of the data presented on page 415 of their article reveals that

on five of nine dimensions, the variances are significantly different ($p < .05$). On these five dimensions, the variance in the ratings given by males is larger than the variance in the ratings given by females. The equivalence of the ratings given by male and female analysts is doubtful.

The data presented by Hackman and Oldham (1974, p. 29) regarding ratings given on each of the JDS job dimensions by supervisors and observers are also amenable to further analysis. The variance of ratings provided by observers is greater on every job dimension than the variance in supervisors' ratings, and significantly different ($p < .05$) on four of the eight dimensions. This suggests that the dispersion in ratings given by observers is greater than the dispersion in supervisors' ratings.

V. DISTANCE

In the literature regarding differences between types of raters, distance formulations are seldom used. Multiple univariate tests of differences in means and variances are related to multivariate tests of the equivalence of centroids and dispersions. However, the few authors who have conducted such univariate tests make no mention of distance.

An approach somewhat akin to distance is mentioned in an intriguing footnote by Cornelius and Lyness (1980):

> The three dependent variables reported in this article were all expressed as Pearson product-moment correlations among ratings. The investigators also collected the same three dependent measures expressed as mean absolute deviations (MABS) in the rating scale values. The results for the MABS measured were largely nonsignificant [p. 158].

The three dependent variables in the Cornelius and Lyness (1980) study were intrarater reliability across time, interrater agreement, and convergence with standard ratings from job analysts and job supervisors. Mean absolute deviation scores is a variant of the distance formula. It appears that the authors preferred using r with its many problems to a type of distance measure (MABS) because significant results were obtained only with r.

VI. SUMMARY

Despite Smith and Hakel's (1979) contention that "who furnishes responses to a job analysis inventory makes little practical difference" (p. 677), the question of the equivalence of job analysis agents remains unanswered. Studies in which correlational approaches have been used provide only tenuous answers because of the many inadequacies inherent in correlation.

The Hackman and Oldham (1974) study in which univariate tests of mean differences were presented and from which univariate tests of differences in variances were performed suggests that different types of agents may not be equivalent. Mean ratings from supervisors, employees and observers differed significantly on five of eight JDS job dimensions. The variance in ratings given by observers was significantly greater than the variance in ratings provided by supervisors on four of the eight dimensions. With respect to the PAQ, Smith and Hakel (1979) reported that mean differences existed among the different types of raters while simultaneously maintaining that the ratings from different job analysis agents are virtually indistinguishable. Nevertheless, evidence

supporting either the equivalence or nonequivalence of job analysis agents is weak at best, and largely nonexistent.

CHAPTER IV

METHOD

I. DESIGN

A balanced one-way multivariate analysis of variance design with four levels of the independent variable (type of rater) and 13 dependent variables (overall factor scores from PAQ) was used to assess the equivalence of different job analysis agents. The four types of raters were job incumbents, their supervisors, students given only the job title, and students who received a description of the job. All four types of analysts rated the job of power-plant operator.

II. SAMPLE AND PROCEDURE

Incumbents and Supervisors

Data came from a large-scale job analysis project recently conducted by a private consulting firm. Part of this project involved incumbents and their supervisors completing the PAQ for the incumbents' jobs. For the job of power-plant operator (Dictionary of Occupational Titles, code 952.382-018), 94 pairs of incumbents and supervisors responded to the PAQ. These pairs of raters came from 30 companies across the country, with one pair of raters per plant and up to six plants per company.

Students

Students in five junior and senior level management classes at The University of Tennessee completed the PAQ as an in-class project. After a short presentation on the importance of job analysis and an introduction to the PAQ, sets of PAQ materials were distributed among the students. Each student received the Position Analysis Questionnaire, the PAQ Record Form (an optically scanned answer sheet), and one of two short lists containing additional instructions which described their specified rating tasks. The only differences between the two forms of instructions was the amount of information provided to the student about the job to be rated. The two forms were randomly distributed within each class.

Both sets of instructions informed the student that:

> The job you will be evaluating is: POWER-PLANT OPERATOR. Other names for this job include POWER ENGINEER and POWERHOUSE ENGINEER. This job is found in light, heat, and power companies.

However, one set of instructions continued with the 183-word description of the job found in the Dictionary of Occupational Titles (p. 923). The form of instructions containing only the job title was received by 94 students; another 94 students received the job description. Both forms are reproduced in the Appendix.

III. MEASURES

PAQ Services, Inc. processed the responses recorded on the PAQ Record Forms. As requested, scores on the 13 overall dimensions of the PAQ were output for each of the 376 participants in the study. These factor scores served as the primary dependent variables.

Dimensions of the PAQ

A factor analysis of the PAQ based on 2,200 jobs produced 13 orthogonal factors. The results of this analysis are reported in depth by Mecham, McCormick and Jeanneret (1977, pp. 104-120). Following is a list of the names of the overall PAQ dimensions and the descriptions provided by Mecham et al. (1977)[1]:

1. Having decision, communicating, and general responsibilities (14% of Total Variance). This dimension is the most inclusive of all the dimensions, having significant correlations with many of the job elements in the PAQ. The dimension reflects activities involving considerable amounts of responsibility for decision making, communicating, and general responsibility. . . .
. . . . . .
2. Operating machines/equipment (9% of Total Variance). This dimension characterizes activities in which individuals are responsible for the operation of machines, equipment, tools, and other types of mechanical and related devices. . . .
. . . . . .
3. Performing clerical/related activities (3% of Total Variance). The dominant feature of this dimension is involvement in the performance of typical clerical, office, and related types of activities. . . .
. . . . . .
4. Performing technical/related activities (2% of Total Variance). This dimension covers a variety of activities that, in general, can be characterized as involvement in the use of various types of technical and related devices, or performing technical types of work without such devices. . . .
. . . . . .
5. Performing service/related activities (2% of Total Variance). The common theme of the job activities covered by this dimension is that associated with performing some services typically also are accompanied by various types of sensory and manual activities. . . .
. . . . . .

---

[1]The numbering of the factors differs from that reported by Mecham et al. (1977). Also, the proportion of variance accounted for by Factor 10 is omitted from the document prepared by Mecham et al. (1977).

6. Other work schedules versus working regular day schedules (2% of Total Variance). The primary distinction represented by this dimension is that of working nontypical day schedules (such as shift work), and irregular kinds of work schedules as opposed to a typical day schedule. . . .

. . . . . .

7. Performing routine/repetitive activities (2% of Total Variance). This dimension is characterized by the performance of routine, repetitive work activities, in some instances at pre-determined work paces. . . .

. . . . . .

8. Being aware of work environment (6% of Total Variance). This dimension typically involves continual awareness of, or sensitiveness to, the environment within which the individual is involved, such awareness being based on the use of various senses, such as vision, hearing, etc. In addition, the dimension typically involves making some kind of response to changing environmental conditions, such as the use of various kinds of control mechanisms, the operation of vehicles, etc. . . .

. . . . . .

9. Engaging in physical activities (3% of Total Variance). The dominant feature of this dimension is involvement in general body or physical activities such as walking, stooping, standing, handling, etc. Implicit in such involvement in some instances is the possibility of physical hazards and associated physical impairment. . . .

. . . . . .

10. Supervising/directing/estimating. This dimension involves several often related coordinating, directing, and estimating functions, frequently but not always associated with supervisory or management positions. . . .
11. Public/customer/related contacts (1% of Total Variance). This dimension is characterized dominantly by the need for personal contacts with the public, customers, or other individuals such as clients, patients, etc. . . .

. . . . . .

12. Working in an unpleasant/hazardous/demanding environment (3% of Total Variance). This dimension is characterized by a spectrum of job environments that usually would be considered as unpleasant, potentially hazardous, or personally demanding. . . .

. . . . . .

13. Non-typical work schedule/optional apparel (1% of Total Variance). This dimension is not as clearly delineated as some of the others in that it is dominated by an admixture of job elements dealing with non-typical work schedules (such as irregular work and night schedules) and dealing with apparel (optimal and informal apparel).

The "opposite" end of the dimension is characterized by more regular work schedules. Aside from being a rather unclear dimension it also is very unimportant since it accounts for only 1% of the total variance.

CHAPTER V

ANALYSES AND RESULTS

Scores on the 13 overall dimensions of the PAQ formed the data base for the statistical analyses which were performed. Table 4 displays the means and variances of the factor scores and the intercorrelations among the factors. Although these 13 factors were uncorrelated "in the normative sample of 2,200 jobs on which these dimensions were derived" (Mecham et al., 1977, p. 104), the presence of significant intercorrelations in the present sample of one job is not surprising. The correlations shown in the table were computed across people, not across jobs. Moreover, even if the normative sample were similar to the sample used here, factor scores often correlate among themselves even when the factors are orthogonal (Gorsuch, 1974, p. 232). Although the factor scores are correlated, the factors retain the conceptual meaning ascribed to them by Mecham et al. (1977).

I. CORRELATIONAL APPROACH TO ASSESSING SIMILARITY

So that a comparison between the results produced by the problem-prone correlational approach and the proposed multivariate strategy can be made, correlations among the different types of raters over the 13 PAQ dimensions were calculated. Figure 7 shows the patterns of the mean ratings for each group across the dimensions. The correlations among these mean profiles are shown in Table 5. Even though the correlations

TABLE 4

CORRELATIONS AMONG OVERALL PAQ DIMENSIONS

Factor	Factor 1	2	3	4	5	6	7	8	9	10	11	12	13	Mean	Variance
1	1.00													.457	.395
2	.16[a]	1.00												1.031	.494
3	.02	.27	1.00											.287	.486
4	-.15	.01	.02	1.00										- .114	.799
5	.18	.31	.00	.27	1.00									- .596	1.357
6	.40	.36	-.12	.06	.58	1.00								-1.564	1.468
7	-.02	-.18	-.10	-.09	-.40	-.23	1.00							- .024	.679
8	.15	.25	.33	.00	.34	.26	-.17	1.00						.063	.214
9	.11	-.30	.10	-.29	-.36	-.18	.10	-.10	1.00					- .732	.529
10	.25	.19	-.01	.02	.23	.25	-.17	.25	-.06	1.00				1.481	.789
11	.28	.33	.03	.06	.58	.37	-.38	.31	-.10	.25	1.00			- .652	1.134
12	.37	.32	.04	-.05	.58	.59	-.33	.23	-.10	.25	.55	1.00		.841	3.252
13	.32	.10	.24	.09	.17	.17	-.39	.16	-.08	.13	.29	.27	1.00	.865	1.062

[a]Critical values for r = .10 for $p < .05$, .13 for $p < .01$, and .17 for $p < .001$; two-tailed tests (N = 376).

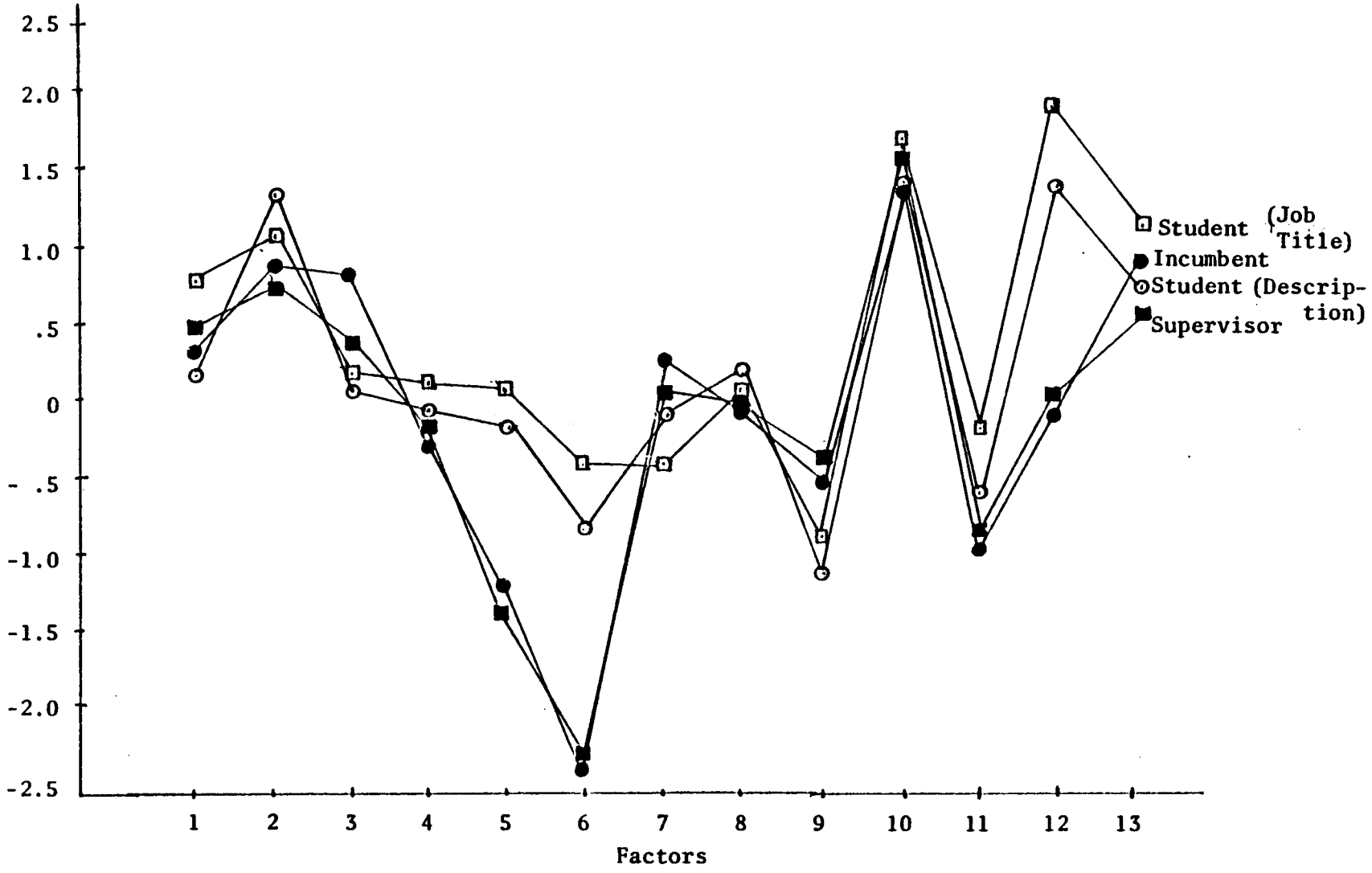


Figure 7. Profiles of group means across overall dimensions of PAQ.

TABLE 5

CORRELATIONS AMONG MEAN PROFILES OF DIFFERENT TYPES OF ANALYSTS

	Type of Analyst			
Analyst	Incumbent	Supervisor	Student (Description)	Student (Job Title)
Incumbent	1.00			
Supervisor	.98**	1.00		
Student (Description)	.67**	.60*	1.00	
Student (Job Title)	.74**	.71**	.94**	1.00

*$p < .05$, two-tailed test.

**$p < .01$, two-tailed test.

are calculated across only 13 dimensions, all reach statistical significance. Of special interest is the high correlation between the mean incumbent ratings and the mean supervisor ratings ($\underline{r}$ = .98).

To further examine the similarity between incumbents and supervisors, a supplemental correlational analysis was performed. A separate correlation was calculated between each incumbent and the incumbent's own supervisor. These 94 correlations ranged from .20 to .98. The median of this distribution of Pearson product-moment correlation coefficients (median $\underline{r}$ = .74) and the mean of the distribution using Fisher Z transformations ($\overline{\underline{r}}$ = .76) are statistically significant ($p < .01$).

Statistical tests for the significance of the difference between the correlation coefficients have little statistical power because of the small number of dimensions over which the correlations were computed. Nevertheless, when Formula 5 (p. 40) was applied to each of the 15 pairwise comparisons which could be made between the correlation coefficients, three of the differences were significant ($p < .01$). Supervisors and incumbents were more highly correlated than were either supervisors and students (job title), supervisors and students (description), or incumbents and students (descriptions).

II. MULTIVARIATE APPROACH TO ASSESSING EQUIVALENCE

The ratings from different types of raters are considered equivalent if and only if the centroids and dispersions of their multivariate distributions are the same. Separate statistical tests were employed to determine the equivalence of the centroids and dispersions of the multivariate distributions of the four types of analysts in the 13

dimensional space. Neither centroids nor dispersions can be significantly different for two or more groups to be considered equivalent.

Equivalence of Dispersions

To measure dispersion, the squared distance between each rater's position in the multidimensional space (determined by the pattern of the 13 factor scores) and the corresponding group centroid was computed by means of the Mahalonobis $\underline{d}^2$ (using the pooled within-group variance-covariance matrix). Members in a group with a relatively large dispersion will generally be located farther from the group's centroid than members in a compact group (i.e., distribution with small dispersion).

The mean or average squared distance from the centroid for the members of a group is a statistic which could logically be used to describe the dispersion of a group. These sample means of the dispersion could then be compared to make inferences concerning the relative dispersions of the ratings given by the four types of raters. However, comparisons of the average value of $\underline{d}^2$ within each group using parametric statistics is deemed inadvisable because of the positive skew in the sample distributions of $\underline{d}^2$. (Skew ranged from 1.3 to 1.6 within the four sample distributions.) Instead, the median $\underline{d}^2$ for the groups of raters could be compared using nonparametric statistical methods which make fewer demands concerning the shape of the sample distributions, the level of measurement and the sample size.

The Kruskal-Wallis procedure, a nonparametric analogue of a one-way ANOVA, is sensitive to differences in medians among two or more groups (Gibbons, 1976). The results of this procedure show that significant

differences exist among the four distributions. The median $\underline{d}^2$ for the groups and the corresponding test statistic are presented in Table 6.

However, the Kruskal-Wallis procedure applied to the four group medians does not identify exactly which groups differ in dispersion. The Kruskal-Wallis test statistic could be computed separately for each possible pairwise comparison of medians, but this approach would lead to an inflated risk of a Type I error. Fortunately, Gibbons (1976, pp. 181-192) described a procedure attributed to Dunn (1964) for making all pairwise tests while maintaining the risk of a Type I error at or below a specified value across all comparisons. Pairs of medians sharing a common superscript in Table 6 are significantly different. Incumbents are significantly less disperse than are either group of students. Further, supervisors are significantly less disperse than are students given only the job title.

To further explore the differences among the groups in terms of the variability of their ratings, univariate analyses were conducted using Cochran's $\underline{C}$ statistic. The variances on each of the 13 dimensions for each group are shown in Table 7. Significant differences in variances existed among the groups on Factors 3, 4, 5, 11, and 12. Because of the large number of pairwise comparisons to be made, a pair of variances was not declared significantly different unless the risk of a Type I error for that comparison was $\leq .01$. Of the 78 pairwise comparisons, 16 were significant. No significant differences between the two student groups were found. Only one of the pairwise contrasts involving incumbents and supervisors was significant; the variance among incumbents

TABLE 6

MEDIAN $\underline{d}^2$ FOR EACH TYPE OF ANALYST

	Type of Analyst				
	Incumbent	Supervisor	Student (Description)	Student (Job Title)	$\underline{X}^2$
Median $\underline{d}^2$	$9.48^{a,b}$	10.36^{c}	11.78^{a}	$14.03^{b,c}$	38.96*

^{a,b,c}Medians with common superscripts are significantly different ($\underline{p} < .05$).

*$\underline{p} < .0001$.

TABLE 7

VARIANCES ON OVERALL DIMENSIONS BY DIFFERENT TYPES OF ANALYSTS

PAQ Dimensions	Type of Analyst				Cochran's C
	Incumbent	Supervisor	Student (Description)	Student (Job Title)	
Factor 1	.24	.39	.40	.38	.29
Factor 2	.42	.54	.36	.43	.31
Factor 3	.36	.56	.39	.42	.32*
Factor 4	$.62^{b}$	$.59^{a}$	.73	$1.15^{a,b}$	.37**
Factor 5	$.64^{c,d}$	$.53^{a,b}$	$1.14^{b,d}$	$1.84^{a,c}$	.44**
Factor 6	$.41^{a,c,d}$	$.72^{a}$	$.74^{d}$	$.90^{c}$	.32
Factor 7	.43	.70	.71	.65	.29
Factor 8	.18	.17	.27	.23	.32
Factor 9	.39	.37	.51	.48	.29
Factor 10	.60	.78	.76	.94	.31
Factor 11	$.57^{b,c}$	$.80^{a}$	1.16^{c}	$1.71^{a,b}$	.40**
Factor 12	$1.64^{c,d}$	$1.43^{a,b}$	$3.29^{b,d}$	$3.73^{a,c}$	.37**
Factor 13	.92	.87	.95	1.38	.34*

^{a,b,c,d}Variances with common superscripts within the same row are significantly different ($p < .01$).

*$p < .05$, two-tailed test.

**$p < .01$, two-tailed test.

was significantly less than the variance among supervisors on Factor 6 (other work schedules versus working regular day schedules).

Equivalence of Centroids

Methods of determining the equivalence of centroids are well known. Multivariate analysis of variance (MANOVA) and discriminant function analysis have been developed for this purpose. In fact, Lissitz et al. (1979) mentioned MANOVA as an approach to a statistically similar problem—that of assessing the equivalence of job centroids.

One of the assumptions of MANOVA is homogeneity of the within-group variance-covariance matrices. After having examined the univariate variances presented in the discussion of dispersion, it is not surprising to find that the within-group variance-covariance matrices are heterogeneous (Box's M = 613.5, $p < .0001$). Fortunately, MANOVA is considerably robust with respect to violations of this assumption when a balanced design (i.e., equal numbers of observations in each group) is used (Hakstian, Roed, & Lind, 1979).

A one-way MANOVA revealed that significant differences exist among the four group centroids (F = 18.6; df = 39,1066; $p < .0001$). Moreover, pairwise F ratios showed that every possible pair of groups are significantly different ($p < .0001$) (see Table 8).

However, statistical significance does not necessarily imply practical significance. A small effect and a large number of degrees of freedom may be statistically significant, but of little actual importance. Indices of strength of association provide insight into the practical importance of a relationship. Lissitz et al. (1979) recommended an

TABLE 8

PAIRWISE SIGNIFICANCE TESTS ON GROUP CENTROIDS: F and $\hat{\omega}^2_m$

	Type of Rater		
Type of Rater	Incumbent	Supervisor	Student (Description)
Supervisor	3.67[a] (.210)[b]		
Student (Description)	36.02 (.727)	36.15 (.727)	
Student (Job Title)	33.76 (.714)	26.52 (.662)	7.40 (.351)

[a]Critical value of F (df = 13,174) = 2.85; p < .001.

[b]Values of $\hat{\omega}^2_m$ corresponding to values of F are shown in parentheses.

index given by Tatsuoka (1970) for calculating the proportion of generalized variance of the dependent vector (in this case, PAQ factors) that can be attributed to the independent variables (type of rater):

$$\hat{\omega}^2_m = 1 - \frac{N\,\lambda}{(N - K) + \lambda}$$

where λ = Wilk's likelihood ratio statistic, N = total sample size, and K = number of types of raters.

The smallest F displayed in Table 8 is that from a MANOVA using only supervisors and incumbents as levels of the independent variable (F = 3.52). To determine if the difference between supervisors and incumbents is of practical significance, F was transformed to λ (Tatsuoka, 1971, p. 89) and $\hat{\omega}^2_m$ was computed. Its value was found to be .21. (Values of $\hat{\omega}^2_m$ corresponding to the pairwise F ratios are shown in Table 8.)

The interpretation of $\hat{\omega}^2_m$ parallels that of the interpretation of R^2 or an univariate $\hat{\omega}^2$. The larger the value of the statistic (maximum = 1.0), the greater is the strength of the relationship or the effect of the independent variable. The extent of differentiation between incumbents and supervisors was found to be .21 (using $\hat{\omega}^2_m$). Although .21 may not seem particularly impressive, it does indicate modest difference when the design of the study is taken into account. The source of data was 94 pairs of supervisors and incumbents in 94 different plants in 30 companies across the country. Although all incumbents were employed in jobs with the same code number in the Dictionary of Occupational Titles, some differences in the actual characteristics of their jobs would be expected.

Differences in geographic location, differences in companies and differences in plants likely produced somewhat different duties for the power-plant operators. These differences would be reflected in ratings of the job on the PAQ. The effect of these differences would be to increase the generalized variance (or dispersion) of the PAQ scores. In computing the multivariate F and $\hat{\omega}^2_{\underline{m}}$, differences in the 94 jobs are treated as error. In spite of this "built-in" error, a full 21 percent of the generalized variance could be explained by the independent variable (supervisor versus incumbent).

The values of $\hat{\omega}^2_{\underline{m}}$ associated with other pairwise contrasts are all at least equal to .35. Further, 78.2 percent of the generalized variance of the PAQ factors can be attributed to type of analyst when all four types of raters are entered into the analysis. The size of these indices of the differentiation among types of raters suggests that the effect of type of analyst is certainly of practical significance as well as being statistically significant.

To better visualize the relative positions of the four group centroids in 13 dimensional space, a discriminant function analysis was performed. All three possible functions were significant ($p < .01$). Attempting to interpret these functions based on the standardized discriminant function coefficients shown in Table 9 is futile because of the conceptual independence of the 13 PAQ factors. Nevertheless, group centroids in the three discriminant function space are specified in Table 10 and the relative positions of the centroids in two discriminant function space is graphically displayed in Figure 8. The first function clearly acts to separate incumbents and supervisors from the two student

TABLE 9

STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS

PAQ Dimensions	Discriminant Function		
	Function I	Function II	Function III
Factor 1	-.37172	.82689	.03251
Factor 2	.29990	-.35385	-.14194
Factor 3	-.44485	-.12228	.70169
Factor 4	.12929	.42377	.18200
Factor 5	-.03013	-.02571	.19471
Factor 6	.93992	.01941	.22685
Factor 7	-.02667	-.63823	.22389
Factor 8	-.15977	.08754	.12945
Factor 9	-.35831	.28791	.01303
Factor 10	-.20521	.12514	-.37553
Factor 11	-.18202	.03939	-.08207
Factor 12	.34031	-.14124	-.11303
Factor 13	.13643	-.15754	.42840

TABLE 10

COORDINATES OF GROUP CENTROIDS IN THREE DISCRIMINANT FUNCTION SPACE

Type of Analyst	Discriminant Function		
	Function I	Function II	Function III
Incumbent	-1.538	-.364	.341
Supervisor	-1.403	.332	-.368
Student (Description)	1.565	-.739	-.149
Student (Job Title)	1.376	.771	.176

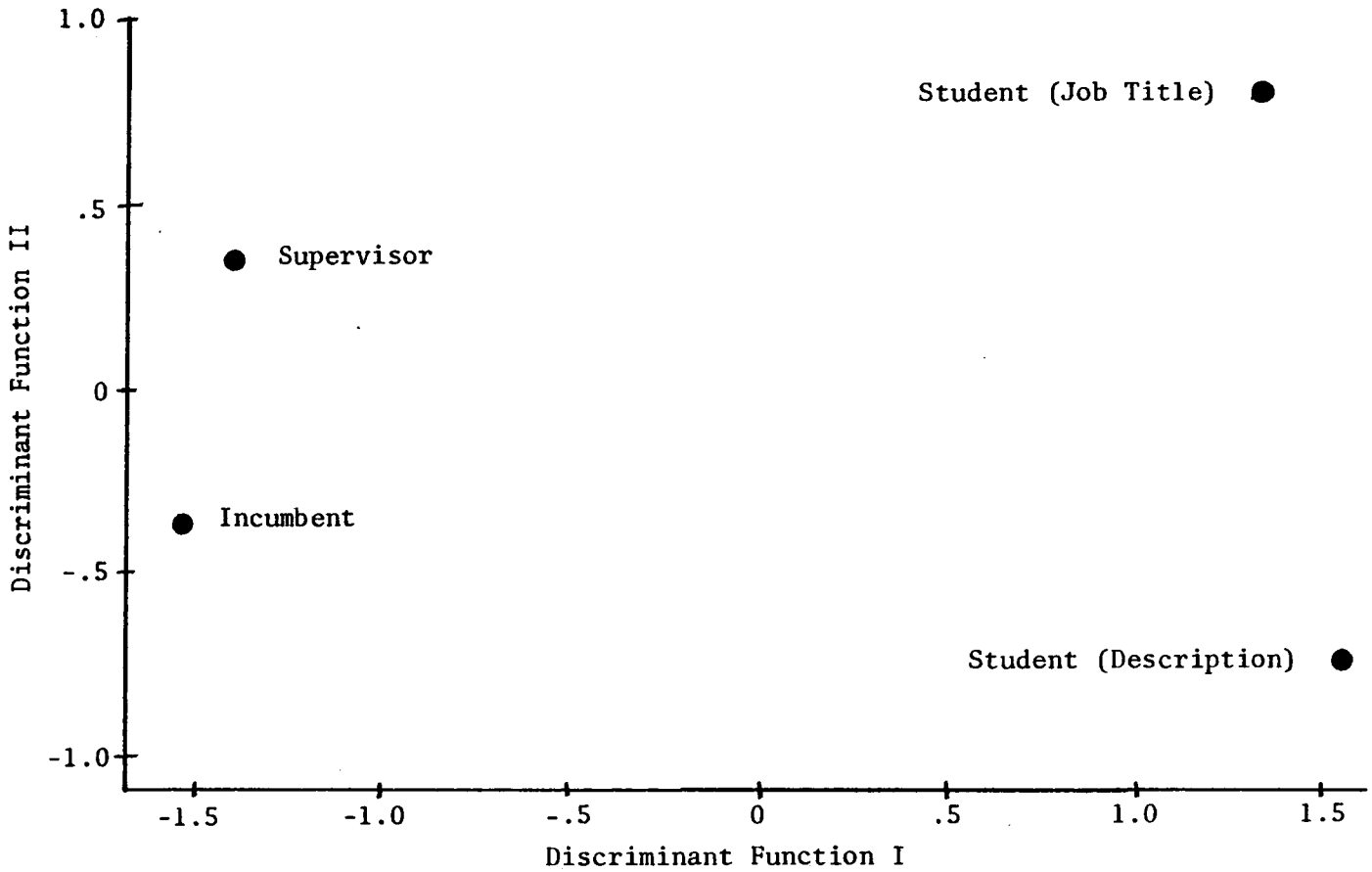


Figure 8. Plot of group centroids in two discriminant function space.

groups. The second function serves to discriminate incumbents from supervisors and to distinguish students given only the job title from those who received the job description. It can be seen from the coordinates of the third function in Table 10 that, like the first function, it serves primarily to separate the student groups from the supervisor and incumbent groups.

To examine differences among the groups at the univariate level, a one-way ANOVA was performed for each of the 13 factors. Table 11 displays the results. Significant differences exist among the groups on 12 of the 13 factors ($p < .01$). Tukey's Honestly Significant Difference (HSD) method is recommended by Myers (1972) for making multiple post hoc pairwise comparisons among means when group sizes are equal. The risk of a Type I error for the six pairwise comparisons performed for each factor is never greater than .05. Of the 78 contrasts between means, exactly half are significant. However, the means for incumbents and supervisors were significantly different on only Factor 3 (performing clerical/ related activities). Incumbents rated the job as requiring more clerical activities than did supervisors.

TABLE 11

MEANS ON OVERALL DIMENSIONS BY DIFFERENT TYPES OF RATERS

PAQ Dimensions	Type of Rater				F*
	Incumbent	Supervisor	Student (Description)	Student (Job Title)	
Factor 1	0.305^{a}	$0.487^{b,c}$	$0.242^{b,d}$	$.796^{a,c,d}$	16.5
Factor 2	$0.862^{a,b}$	$0.730^{c,d}$	$1.362^{a,c}$	$1.169^{b,d}$	17.9
Factor 3	$0.688^{a,b,c}$	0.261^{a}	0.052^{b}	0.149^{c}	17.2
Factor 4	-0.263^{a}	-0.246^{b}	-0.131	$0.184^{a,b}$	5.2
Factor 5	$-1.170^{a,b}$	$-1.139^{c,d}$	$-0.212^{a,c}$	$0.139^{b,d}$	39.6
Factor 6	$-2.478^{a,b}$	$-2.354^{c,d}$	$-0.871^{a,c,e}$	$-0.480^{b,d,e}$	140.8
Factor 7	0.257^{a}	0.038^{b}	0.051^{c}	$-0.443^{a,b,c}$	13.4
Factor 8	0.080	-0.001	0.024	0.150	2.0
Factor 9	$-0.504^{a,b}$	$-0.390^{c,d}$	$-1.165^{a,c,e}$	$-0.868^{b,d,e}$	26.9
Factor 10	1.258^{a}	1.584	1.409	1.674^{a}	4.2
Factor 11	-0.975^{a}	-0.823^{b}	-0.596	$-0.213^{a,b}$	9.7
Factor 12	$-0.092^{a,b}$	$0.078^{c,d}$	$1.476^{a,c}$	$1.904^{b,d}$	37.0
Factor 13	0.870	0.664^{a}	0.743^{b}	$1.182^{a,b}$	4.7

^{a,b,c,d,e}Means with common superscripts within the same row are significantly different ($p < .05$).

*Critical value of F ($df = 3,372$) = 3.84; $p < .01$.

CHAPTER VI

SUMMARY AND IMPLICATIONS

I. SUMMARY

Overview

The purpose of this research was to conceptually and empirically investigate the equivalence of ratings provided by individuals with different relations to a job. Specifically, ratings of the job of power-plant operator given by job incumbents, their supervisors, students with access to the job description, and students who were only told the job title were examined for equivalence.

Several existing methods for assessing similarity of ratings were reviewed before the empirical investigation was undertaken. The Pearson r and the intraclass r were shown to be dependent on the direction of measurement and to be limited to a finite range of possible values. Another deficiency in the Pearson r for assessing similarity is its inability to detect differences in elevation. However, distance measures such as the Mahalonobis d^2 suffer from none of these inadequacies.

A definition of rater equivalence was developed which involved the use of distance measurements in multivariate space. By this definition, the ratings from different types of raters are considered equivalent if and only if both the centroids and the dispersions of their multivariate distributions are the same. Equivalence of centroids assures that significant differences do not exist between types of raters in the mean

ratings of a job on different dimensions. Equivalence of dispersions guarantees that interrater agreements within groups of different types of analysts do not differ significantly. Relatively little dispersion among raters of a job is associated with relatively greater statistical power for detecting differences between jobs using either a univariate approach (e.g., Arvey & Mossholder, 1977) or a multivariate strategy (e.g., Lissitz et al., 1979).

Both the correlational approach to measuring similarity among raters and a set of multivariate techniques developed from the definition of rater equivalence were applied to the same set of sample data. However, before reviewing the findings of the research, limits to the generalizability of the results must be noted.

Limitations

The most obvious limitation concerns sampling of the job which was analyzed. The reported results correspond to the ratings given to only one job (viz., power-plant operator). Further, this job was not selected at random, but rather was chosen because of the availability of ratings from pairs of supervisors and incumbents on the job. The possibility exists that findings based on the job of power-plant operator may not generalize beyond this one job, although reasons for such a limitation are not apparent.

The other important limitations concern the method of job analysis and the types of analysts. The results of this study may not generalize to other job analysis methods or even other structured job analysis questionnaires. The results could be specific to the Position Analysis

Questionnaire. Also, conspicuous by its absence is data collected from job analysts, one of the three types of job analysis agents specified by McCormick (1976). The relatively large sample sizes needed for multivariate analyses and the practical difficulties involved in locating a sufficiently large sample of job analysts prevented the inclusion of job analysts in this research. Thus, the equivalence of ratings provided by job analysts and the ratings given by job incumbents and their supervisors cannot be ascertained in the present study.

Findings

The correlational analyses produced pairwise correlation coefficients between types of raters across mean ratings on the 13 overall PAQ dimensions ranging from .60 to .98. The correlations computed in a similar manner across the mean ratings of 187 PAQ items in the Smith and Hakel (1979) study ranged from .89 to .98. The correlations involving students are higher in the Smith and Hakel (1979) research ($\bar{r}$ = .93) than are the five correlations involving students shown in Table 5 (p. 65) which were found in this study ($\bar{r}$ = .76). The difference between these two average correlation coefficients is statistically significant ($p < .05$). Perhaps students' unfamiliarity with the job of power-plant operator reduced the size of the obtained correlations. However, in both studies the correlation between supervisors and incumbents was .98. A separate correlation for each of the 94 pairs of supervisors and incumbents across the 13 dimensions was also computed. The average correlation coefficient ($\bar{r}$ = .76) is very similar to the average r between supervisors and incumbents which McCormick et al. (1977) reported ($\bar{r}$ = .79).

Results of the correlational analyses revealed a high degree of correspondence among the four types of raters, particularly between supervisors and incumbents. The size of the obtained correlation coefficients is similar to that reported by previous researchers. The significant correlations among the raters seem to provide support for Smith and Hakel's (1979) contention that the source of job analysis information makes little difference.

Following the correlation analyses, data were subjected to multivariate tests to determine the equivalence of different raters. Significant differences were found among the groups both with respect to dispersions and centroids. None of the four types of raters met the requirements for being labeled equivalent to any other type.

Differences in dispersions indicated differences in interrater agreement. Incumbents displayed most agreement among themselves and students given the job title showed least agreement. Supervisors were less disperse than were students who read the job description. However, after groups were ranked in terms of dispersion, those groups which were next to each other were not found to be significantly different.

The order of the size of the dispersions seems to correspond to the amount of information known about the job. The closer an individual is to the job, the more about the job the individual should know. A person actually performing the job might be expected to know more about the nature of that job than someone supervising the position. Ratings from students which are based on scanty information are likely to be unreliable and open to challenge. More knowledge about the job should lead to more precise responses to the PAQ and greater interrater agreement. If this

speculation concerning the order of the dispersions is correct, it suggests that the validity of job analysis ratings is a function of the rater's closeness to the job. Although this idea may be worth pursuing in future research, the issue of validity was not addressed in this study.

The MANOVA revealed highly significant differences among the four group centroids. When all four types of analysts form the levels of the independent variable, 78.2 percent of the generalized variance of the PAQ factors can be explained by the independent variable (computed with $\hat{\omega}^2_m$). This strong differentiation among types of analysts demonstrates the practical significance of the statistically significant multivariate F.

The follow-up univariate F ratios computed for the individual PAQ dimensions showed that the four groups differed on 12 of the 13 dimensions. Unlike the findings concerning dispersions and the MANOVA of the centroids, these findings have some counterparts in previous research. Smith and Hakel (1979) found significant differences among types of analysts in their mean ratings on various PAQ items. Hackman and Oldham (1974) reported mean differences among different types of raters on several dimensions of the Job Descriptive Survey.

From a practitioner's point of view, the most important findings relate to the equivalence of job incumbents and their supervisors in the rating of a job. As already mentioned, the two groups did not differ significantly in dispersions. However, a MANOVA on the two group centroids was highly significant ($p < .0001$), demonstrating that the two groups are not equivalent. The value of $\hat{\omega}^2_m$ (.21) indicated that the degree of differentiation between the two groups is modest.

The correlational analyses and the multivariate analyses conducted on the same set of data provide results which are not ambiguous but are contradictory. Significant correlations indicated that the types of raters were similar, whereas significance tests from the multivariate analyses showed that the types of raters were different. A very different set of conclusions is reached, depending on which statistical method is used. The correlational analyses strongly support Smith and Hakel's (1979) contention that incumbents and supervisors provide equivalent responses to a job analysis inventory. The support is less strong, but still convincing, that students provide ratings which correlate substantially with the more traditional types of job analysis agents. However, the analysis of centroids and dispersions clearly demonstrated that ratings from different types of raters cannot be considered equivalent and that researchers may be misled if they rely on correlations to assess similarity.

II. IMPLICATIONS

Rating the Job

When collecting job analysis information from supervisors and incumbents, differences should be anticipated. Differences among the raters could be resolved by using various group decision making strategies such as the consensus approach (Hall, 1971). However, groups often reach agreement in such situations by merely averaging the ratings given by individuals (e.g., Gordon & Isenberg, 1975). If the ratings are to be averaged, the time involved in group process can be eliminated by having a computer calculate the average rating on each dimension for

the job. So that ratings of different jobs are comparable, the same mix of supervisors and incumbents should rate each job. This suggestion is consistent with Lawshe's (1975) prescription that equal numbers of incumbents and supervisors be involved in generating content validity information.

If the most important criterion in determining who should provide job analysis data is high interrater agreement, the results of this study provide practical guidelines. Either incumbents or supervisors should rate a job, but not incumbents and supervisors. Interrater agreement (as measured by dispersion) is much less if the ratings of the two groups are combined into one group. That is, the dispersion among incumbents and the dispersion among supervisors are less than the dispersion in a system in which the two types of raters are combined.

Usefulness of PAQ

The finding that students' ratings are very different than those marked by supervisors and incumbents supports the usefulness of the PAQ. A person relatively unfamiliar with a job (e.g., student given only a job title) should rate the job very differently than someone with intimate knowledge of the job. Without an understanding of the nature of the work, an individual completing the PAQ must often guess which is the "correct" response. If the processes involved in rating a job in job analysis and the evaluation of an individual in performance appraisal are similar, then the study by Landy and Guion (1970) is relevant to this discussion. They found that the opportunity to observe job relevant behavior of a person being evaluated affects the reliability of the

ratings. However, Smith and Hakel (1979) apparently disagree with the viewpoint that how much a person knows about a job is related to the job analysis information which can be provided.

On the basis of their correlational analyses which showed marked similarity among different raters, Smith and Hakel (1979) stated:

> These findings also seem to raise some questions about the sensitivity of the PAQ and other structured job analysis questionnaires to differentially diagnose job information. At first glance it seems that incumbents, supervisors, and analysts should produce job analysis data that is markedly different from that which is produced by people unacquainted with the work (such as students given only a job title). . . . If one's purpose is job evaluation, job classification, or the group of similar occupations into clusters for test validation purposes, then the fact that even lay persons with little contact with the job can agree at extremely high levels with incumbents, supervisors, and analysts should be interpreted as evidence for the usefulness of the instrument. The level of convergence reported here is evidence of widespread knowledge about the job duties [pp. 690-691].

The logic of this argument is difficult to understand. Smith and Hakel (1979) provide little rationale or explanation for their conclusion concerning what constitutes evidence for the usefulness of the PAQ. If their conclusion is accurate, then the findings of the present research raise serious questions concerning the usefulness of the PAQ for job evaluation, job classification, and the grouping of jobs into job families. However, what Smith and Hakel (1979) observed "at first glance) appears to be an accurate insight into "the sensitivity of the PAQ." The sensitivity of a structured job analysis questionnaire should be questioned if persons not familiar with the job produce job analysis data which are not different than that from incumbents and supervisors. Results of the present study provide evidence for the sensitivity and usefulness of the PAQ.

Generalizability of Method

The operationalization of the definition of rater equivalence need not be limited to the domain of job analysis or to McCormick's (1976) types of job analysis agents. Content validity ratings and performance appraisal judgments could similarly be analyzed to assess the equivalence of different types of raters who differ on individual characteristics (e.g., sex, tenure, or educational level). For example, the equivalence of multidimensional performance ratings given by women managers to those judgments recorded by their male counterparts in an organization could be assessed (cf., Rosen & Jerdee, 1973) by applying the multivariate statistical procedures developed from the definition of equivalence. Perhaps researchers in other fields (such as social psychology) can find applications for the described multivariate techniques in their own areas of specialization.

LIST OF REFERENCES

LIST OF REFERENCES

Arvey, R. D., & Mossholder, K. M. A proposed methodology for determining similarities and differences among jobs. Personnel Psychology, 1977, 30, 363-374.

Arvey, R. D., Passino, E. M., & Lounsbury, J. W. Job analysis results as influenced by sex of incumbent and sex of analyst. Journal of Applied Psychology, 1977, 62 (4), 411-416.

Ash, P. The reliability of job evaluation rankings. Journal of Applied Psychology, 1948, 32, 313-320.

Block, J. Amendation to Cohen's cautions regarding the measurement of profile similarity. Psychological Bulletin, 1970, 73, 307-308.

Borman, W. C. Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 1977, 20, 238-252.

Burt, C. The factors of the mind. New York: Macmillan, 1941.

Cascio, W. F. Applied psychology in personnel management. Reston, VA: Reston Publishing, 1978.

Cattell, R. B. r_p and other coefficients of pattern similarity. Psychometrika, 1949, 14, 279-298.

Chou, Y. Statistical analysis with business and economic applications (2nd ed.). New York: Holt, Rinehart and Winston, 1975.

Cohen, J. r_c: A profile similarity coefficient invariant over variable reflection. Psychological Bulletin, 1969, 71, 281-284.

Coombs, C. H., Dawes, R. M., & Tversky, A. Mathematical psychology: An elementary introduction. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Cornelius, E. T., Carron, T. J., & Collins, M. N. Job analysis models and job classification. Personnel Psychology, 1979, 32, 693-708.

Cornelius, E. T., & Lyness, K. S. A comparison of holistic and decomposed judgment strategies in job analysis by job incumbents. Journal of Applied Psychology, 1980, 65 (2), 155-163.

Cronbach, L. J. Proposals leading to analytic treatment of social perception scores. In R. Tagiuri and L. Petrullo (Eds.), Person perception and interpersonal behavior. Stanford, CA: Stanford University Press, 1958.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

duMas, F. M. A quick method of analyzing the similarity of profiles. Journal of Clinical Psychology, 1946, 2, 80-83.

duMas, F. M. On the interpretation of personality profiles. Journal of Clinical Psychology, 1947, 3, 57-65.

duMas, F. M. The coefficient of profile similarity. Journal of Clinical Psychology, 1949, 5, 123-131.

duMas, F. M. A note on the coefficient of profile similarity. Journal of Clinical Psychology, 1950, 6, 300-301.

Dunn, O. J. Multiple comparisons using rank sums. Technometrics, 1964, 6, 241-252.

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

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Edwards, K. H. R. Note on the significance of correlation coefficients from different samples. Journal of Applied Psychology, 1934, 18, 528-535.

Gibbons, J. D. Nonparametric methods for quantitative analysis. New York: Holt, Rinehart and Winston, 1976.

Gordon, M. E. & Isenberg, J. F. Validation of an experimental training criterion for machinists. Journal of Industrial Teacher Education, 1975, 12, 72-78.

Greenwood, J. M., & McNamara, W. J. Interrater reliability in situational tests. Journal of Applied Psychology, 1967, 51, 101-106.

Gregson, R. A. M. Psychometrics of similarity. New York: Academic Press, 1975.

Guilford, J. P., & Fruchter, B. Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill, 1978.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hackman, J. R., & Oldham, G. R. The job diagnostic survey: An instrument for the diagnosis of jobs and the evaluation of job redesign projects. (Yale University Department of Administrative Sciences Technical Report No. 4.) Yale University, May 1974.

Haggard, E. A. Intraclass correlation and the analysis of variance. New York: Dryden Press, 1958.

Hakstian, A. R., Roed, J. C., & Lind, J. C. Two sample T^2 procedures and the assumption of homogeneous covariance matrices. Psychological Bulletin, 1979, 86 (6), 1255-1263.

Hall, J. Decisions, decisions, decisions. Psychology Today, November 1971, p. 51ff.

Harding, F. D., Madden, J. M., & Colson, K. Analysis of a job evaluation system. Journal of Applied Psychology, 1960, 5, 354-357.

Hazel, J. T., Madden, J. M., & Christal, R. E. Agreement between worker-supervisor descriptions of the worker's job. Journal of Industrial Psychology, 1964, 2, 71-79.

Heerman, E. F. Comments on Overall's "Multivariate methods for profile analysis." Psychological Bulletin, 1965, 63, 128.

Heneman, H. G. Comparison of self- and superior ratings of managerial performance. Journal of Applied Psychology, 1974, 59, 638-642.

Horn, J. L. Significance tests for use with r_p and related profile statistics. Educational and Psychological Measurement, 1961, 21, 363-370.

Howard, K. I., & Diesenhaus, H. I. Direction of measurement and profile similarity. Multivariate Behavioral Research, 1967, 2, 225-237.

Landy, F. J., & Guion, R. M. Development of scales for measurement of work motivation. Organizational Behavior and Human Performance, 1970, 5, 93-103.

Landy, F. J., & Trumbo, D. A. Psychology of work behavior. Homewood, IL: Dorsey Press, 1976.

Lawshe, C. H. A quantitative approach to content validity. Personnel Psychology, 1975, 28, 563-575.

Lawshe, C. H., & Farbo, P. C. Studies in job evaluation: 8. The reliability of an abbreviated job evaluation system. Journal of Applied Psychology, 1949, 33, 158-166.

Lissitz, R. W., Mendoza, J. L., Huberty, C. J., & Markos, H. V. Some further ideas on a methodology for determining job similarities/differences. Personnel Psychology, 1979, 32, 517-528.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Madden, J. M. A comparison of three methods of rating-scale construction. Journal of Industrial Psychology, 1964, 2, 43-50.

McCormick, E. J. Job and task analysis. In M. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the position analysis questionnaire (PAQ). Journal of Applied Psychology, 1972, 56, 347-368.

McCormick, E. J., Mecham, R. C., & Jeanneret, P. R. Position analysis questionnaire technical report (System II). Lafayette, Indiana: Purdue University Bookstore, 1977.

McCormick, E. J., & Tiffin, J. Industrial psychology (6th ed.). Englewood Cliffs, NJ: Prentice-Hall, 1974.

Mecham, R. C., McCormick, E. J., & Jeanneret, P. R. Position analysis questionnaire users manual (System II). Logan, Utah: PAQ Services, Inc., 1977.

Myers, J. L. Fundamentals of experimental design (2nd ed.). Boston: Allyn and Bacon, 1972.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Osgood, C. E., & Suci, G. J. A measure of relation determined by both mean difference and profile information. Psychological Bulletin, 1952, 49, 251-262.

Overall, J. E. Note on multivariate methods for profile analysis. Psychological Bulletin, 1964, 61, 195-198.

Pearson, K. The application of the coefficient of racial likeness to test the character of samples. Biometrika, 1928, 20, 294-300.

Pearson, K. Note on standardisation of method of using the coefficient of racial likeness. Biometrika, 1928, 20, 376-378.

Probst, J. B. An experimental test in evaluating the most essential qualifications for the position of "office clerk." Journal of Applied Psychology, 1932, 16, 644-649.

Rosen, B., & Jerdee, T. H. The influence of sex role stereotypes on evaluations of male and female supervisory behavior. Journal of Applied Psychology, 1973, 57, 44-54.

Shrout, P. E., & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 1979, 86 (2), 420-428.

Sjöberg, L., & Holley, J. W. A measure of similarity between individuals when scoring directions of variables are arbitrary. Multivariate Behavioral Research, 1967, 2, 377-384.

Smith, J. E., & Hakel, M. D. Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. Personnel Psychology, 1979, 32, 677-693.

Stephenson, W. Correlating persons instead of tests. Character and Personality, 1935, 4, 17-24.

Stephenson, W. A new application of correlation to averages. British Journal of Educational Psychology, 1936, 6, 43-56.

Stephenson, W. The inverted factor technique. British Journal of Psychology, 1936, 26, 344-361.

Stephenson, W. Some observations on Q technique. Psychological Bulletin, 1952, 49, 483-498.

Tatsuoka, M. M. Discriminant analysis: The study of group differences. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.

Tatsuoka, M. M. Multivariate analysis: Techniques for educational and psychological research. New York: Wiley & Sons, 1971.

Taylor, L. R. Empirically derived job families as a foundation for the study of validity generalization. Study I. The construction of job families based on the component and overall dimensions of the PAQ. Personnel Psychology, 1978, 31, 325-340.

Taylor, L. R., & Colbert, G. A. Empirically derived job families as a foundation for the study of validity generalization. Study II. The construction of job families based on company-specific PAQ job dimensions. Personnel Psychology, 1978, 31, 341-353.

Tellegen, A. Direction of measurement: A source of misinterpretation. Psychological Bulletin, 1965, 63, 233-243.

Tornow, W. W., & Pinto, P. R. The development of a managerial job taxonomy: A system for describing, classifying, and evaluating executive positions. Journal of Applied Psychology, 1976, 61 410-418.

U.S. Employment Service. Dictionary of occupational titles (4th ed.). Washington, D.C.: U.S. Government Printing Office, 1977.

Webster, H. A note on profile similarity. Psychological Bulletin, 1952, 49, 538-539.

Wiley, L., & Jenkins, W. S. Method for measuring bias in raters who estimate job qualifications. Journal of Industrial Psychology, 1963, 1, 16-32.

APPENDIX

YOUR NAME_____

Course _____

Instructor_____

This study is designed to assess various characteristics of a widely used job analysis instrument—the Position Analysis Questionnaire (PAQ). Your task is to complete the PAQ as best you can given the information printed below.

Follow these steps:

1. Check to see that you have the PAQ Booklet and the scan-form answer sheet. PLEASE MAKE NO MARKS ON THE PAQ BOOKLET—IT WILL BE USED AGAIN!

2. The job you will be evaluating is: POWER-PLANT OPERATOR. Other names for this job include POWER ENGINEER and POWERHOUSE ENGINEER. This job is found in light, heat, and power companies.

3. Based on what you know about this job, complete the PAQ:

   a. Turn the scan-form so that "STEP 3: ENTER RESPONSES TO PAQ ITEMS" is face up.

   b. Read the instructions on pages 1 and 2 of the PAQ Booklet.

   c. Use the scan-form to record your responses to the items beginning on page 4 of the PAQ Booklet. (Omit the items printed on the back cover of the booklet.)

   d. When you have finished all 187 items, turn in the scan-form, the PAQ Booklet, and these instructions. (In order to receive proper credit, make sure that your name appears at the top of this page.)

YOUR NAME_____

Course _____

Instructor_____

This study is designed to assess various characteristics of a widely used job analysis instrument—the Position Analysis Questionnaire (PAQ). Your task is to complete the PAQ as best you can given the information printed below.

Follow these steps:

1. Check to see that you have the PAQ Booklet and the scan-form answer sheet. PLEASE MAKE NO MARKS ON THE PAQ Booklet—IT WILL BE USED AGAIN!

2. The job you will be evaluating is: POWER-PLANT OPERATOR. Other names for this job include POWER ENGINEER and POWERHOUSE ENGINEER. This job is found in light, heat, and power companies.

   Here is a description of the job. Please read it carefully.

> Operates boilers, turbines, generators, and auxiliary equipment at generating plant to produce electricity: Monitors control board and regulates equipment following recording and indicating instruments. Adjusts controls of water and cold feed systems, blowers, and igniters to start up or shut down boilers. Controls operation of boiler auxiliary equipment, such as water and vacuum pumps, coal driers and pulverizers, steam condensers, and soot blowers to insure efficient operation of boilers. Adjusts boiler controls to provide steam at specified temperature and pressure for turbine loads according to power demands. Adjusts controls to regulate speed, voltage, and phase of incoming turbines to coincide with voltage and phase of power being generated. Synchronizes incoming generating units with units in operation and closes circuit breaker at exact instant of coincidence. Monitors gages to determine effect of generator load on related equipment, such as *bus bars* and voltage regulators. Adjusts transformer controls to regulate flow of power between generating stations and substations. Operates switchgear to regulate and transfer power loads to protect maintenance workers engaged in repairing or cleaning equipment. Records malfunctions of equipment, instruments, or controls on logsheet.

3. Based on what you know about this job and the above description, complete the PAQ:

   a. Turn the scan-form so that "STEP 3: ENTER RESPONSES TO PAQ ITEMS" is face up.

   b. Read the instructions on pages 1 and 2 of the PAQ Booklet.

c. Use the scan-form to record your responses to the items beginning on page 4 of the PAQ Booklet. (Omit the items printed on the back cover of the booklet.)

d. When you have finished all 187 items, turn in the scan-form, the PAQ Booklet, and these instructions. (In order to receive proper credit, make sure that your name appears at the top of this page.)

VITA

Robert Earl Burt was born March 19, 1953, in Wilmington, Illinois. Ten years later he moved to Independence, Missouri. In 1971, he graduated at the top of his class from William Chrisman High School. For the next four years he attended the University of Missouri at Columbia on academic and music scholarships. The Bachelor of Arts degree in premedical sciences was awarded to him in 1975.

He entered the graduate program in industrial and organizational psychology at The University of Tennessee at Knoxville in 1976. For the first two years, he served as a graduate research assistant in the Psychology Department and the Management Department. During the last six months of 1978, he accepted two personnel research internships with IBM Corporate Headquarters (White Plains, New York) and IRS National Office (Washington, D.C.). When he returned to the university, he began what were to be two rewarding years teaching undergraduate general management and personnel management courses. He received the Doctor of Philosophy degree with a major in industrial and organizational psychology in December 1980.

The author was a recipient of the Walter Melville Bonham Dissertation Award and holds memberships in Sigma Xi Scientific Research Society of North America, Phi Kappa Phi National Honor Society, and the American Psychological Association.