DOCUMENT RESUME

ED 023 166 EA 001 636

By-Ronan, William W.; Prien, Erich P.
Toward A Criterion Theory: A Review and Analysis of Research and Opinion.
Richardson Foundation, Greensboro, N.C. Creativity Research Inst.
Pub Date Jun 66
Note-110p.
EDRS Price MF-$0.50 HC-$5.60
Descriptors-Behavior, Bibliographies, Evaluation Criteria, *Literature Reviews, *Measurement, *Performance Criteria, *Psychology, *Task Performance

Literature dealing with the development and utilization of work performance criteria is reviewed in terms of (1) the reliability of job performance as a criteria, (2) the reliability of job performance observation as a criteria, (3) the dimensionality of job performance, and (4) extra-individual conditions which modify job performance From the review, theorems and corollaries are formulated, testable hypotheses are derived, and 15 areas in which further research would be useful are suggested. It is concluded that variation in job performance is a result of a wide range of causal influences and that its measurement is nebulous. A 226-item bibliography is included. (HW)

TOWARD A CRITERION THEORY

A REVIEW AND ANALYSIS OF RESEARCH AND OPINION

By

William W. Ronan
Georgia Institute of Technology

and

Erich P. Prien
University of Akron

Published By

The Creativity Research Institute

Of The

Richardson Foundation, Inc.

June, 1966

TOWARD A CRITERION THEORY

A REVIEW AND ANALYSIS OF RESEARCH AND OPINION

By

William W. Ronan

and

Erich P. Prien

Published By

The Creativity Research Institute

Of The

Richardson Foundation, Inc.

June, 1966

Toward a Criterion Theory:
A Review and Analysis of Research and Opinion

William W. Ronan
Georgia Institute of Technology
and Erich P. Prien
University of Akron

A literature review dealing with the development and utilization of work performance criteria has revealed some basic questions concerning criteria. They are: (1) Is job performance reliable? (2) Is observation of job performance reliable? (3) Is job performance unidimensional? (4) Is job performance modified by extra-individual conditions?

Generally a paucity of research information exists in all the areas enumerated above. For example, fewer than 25 studies have investigated directly the important concept of performance reliability.

It is suggested that enough information is available to formulate theorems and corollaries and to derive testable hypotheses. In the concluding section 15 areas of required research are suggested as fruitful for providing needed answers to the questions posed.

The "criterion problem" pervades all areas of psychology. In its most basic form, a criterion is an assumed perfect and true measure of variability, whether that variability is of human behavior or some aspect of group or organizational functioning. For the most part, psychologists have been concerned with variation which is more or less directly related

to individual differences within specific situations or with reference to the particular pattern of experimentally controlled variables. However, the ligitimate scope of criterion investigation includes development of concepts of personality characteristics, characteristics of group and organizational functioning. Investigation is also justified of the more practical problems such as the definition of human emotional adjustment, dimensions of executive performance, dimensions of employee job withdrawal behavior, or the definition of sales performance.

The concern of psychologists and others in research and practice has been with the more practical matters of development and measurement within specific situations. In criterion research, unlike learning theory or personality theory, very little has been done in the area of individual-situation interaction which would qualify as basic or pure research aimed at the development of a theoretical structure. Certainly under the broad scope of the definition, personality theory and theories of social interaction come close to satisfying this void. However, it is seldom that any effort is made to bridge the gap between the study of the individual or group in artificial situations, and variability of behavior and performance in the world of reality.

Much of the empirical work in the various areas of personnel psychology has been a matter of expedience, motivated by the need for solution to a specific problem rather than by the desire to generate a theoretical framework.

Historically, the emphasis has been on the selection of the "most noticeable" rather than on the development of the most appropriate criterion. The tendency has been to accept what existed rather than to determine both the "necessary" and "sufficient" standards. Otis (1953) succinctly identifies the researcher and the practitioner as the culprits in this respect.

Considerable empirical data have been amassed, but there have been few attempts to assess these data in total. A complete survey of the literature in all areas of psychology is, of course, prohibited. Admittedly, the need for criterion research is as present and pressing in other areas of psychology as it is in personnel and industrial psychology. To the extent that other areas of psychology overlap with personnel and industrial psychology some reference will be made to existing empirical data in those areas. However this review is primarily concerned with the problems of variability of performance behavior in work situations. The emphasis is on more objective performance measures with material on merit rating included only to clarify specific points.

By our definition, this review is concerned with behaviors which are limited by operations within specific situations; operational definitions of behavior variability of individual-situation interactions or group-situation interactions. Ultimately the combinations of individual/situational factors should lead to definitions of variability within complete organizations. The ultimate practical solution is the identification of the antecedent conditions, both the individual

differences and the situational characteristics, which limit. enhance or inhibit behavior variability. Ultimately we must understand performance within this context of individual. situational and organizational variables acting separately and interacting to affect performance behavior. Our criterion definition thus is measurements of the manifestations of performance behavior based upon characteristics of individuals as they affect and are affected by situational and organizational characteristics.

Industrial psychology has for many years studied a few of the possible methods for measuring criteria of job performance. The result has been the rather wry cliche. "the criterion problem." A recent statement of this problem was by Dudek (1963) in the Annual Review of Psychology. i. e., "Criterion problems. as usual. received a great deal of attention--and some action." An earlier statement by Viteles (1926) was. "...it requires only a brief survey of the literature to show that in spite of the recognized importance of reliable standards and/or recognized precautions in the selection of such standards, the criteria in individual investigations have on the whole been very unsatisfactory." Essentially the same statement is made by Wallace and Weitz (1955) and Baire (1959) in writing of major findings or problems in industrial psychology. No writer. though. suggests the probability of isolation of the problem (if it is a problem. Dunnette 1963a) in the near future. In general. it appears that attention but little action will continue to be the role.

Historical Overview

As might be inferred from Viteles' quotation, attention had been devoted to the development of adequate measures of job performance for several years. Link (1919) published one of the earliest studies wherein ratings of job performance from two supervisors were secured. Thorndike (1920), based upon earlier work by Wells (1907), named the "halo effect" that to a large degree accounted for the high correlations Link obtained, i.e., .82 and .92, in two different groups. Freyd (1923-24) in a general discussion of vocational selection problems discussed the need for job analysis, the importance of individual differences, the concept of recognizing that different jobs require different abilities and that measurement in these areas was possible. Twelve possible criteria were named and discussed. Investigations using more objective criteria than ratings had begun earlier. Yerkes (1921) presented what appears to be the earliest study using more objective criteria. The criteria were output and accuracy of graphotype operators with a correlation of .11 between the two. Lovett (1923) published a study on selection of salesmen that was very sophisticated for the time and can still be regarded as the exceptional design. Similarly Hornhauser (1923-24) presented a selection study of billing machine operators using eight tests and years of schooling to predict six criteria. This study too had estimates of reliability and intercorrelations of selected criteria. Pond (1925-26) presented another of the earlier studies that, along with the selection basis, made a systematic

study of the reliability and interrelationships of criteria of job performance. In addition to reliability indices of four criteria, Pond intercorrelated foremen's ratings with highest weekly pay. The intercorrelations were of a nature that has become quite well established since this pioneering study, i.e., a range from the -.30's to .50's with a median in the 20's. Her solution was one that has also become all too common -- "These sources of unreliability in the factory criteria of success were themselves unmeasured, and difficult to evaluate in any way. There was always the possibility that in spite of them, significant relationships might be found between the criteria of success and test scores." Concurrently with Pond's work, Shellow (1925-26a) was facing the same problems in studying the selection of street car motormen. She discussed alternative criteria and, in view of disappointing reliabilities (intercorrelation .05 between ratings by the "Chief Instructor" and "member of Educational Department") finally decided upon turnover as a criterion. Another early study by Frey (1925-26) discovered a unique source of criterion bias. "The sales record itself was found to be anerratic measure of sales ability because some of the men ran up high records by selling only to relatives, whereas others of considerable past experience or apparent aptitude lacked temporarily a clientele. The sales managers were able to detect the cases where the sales record was not a valid criterion and make the necessary adjustments." The "rebate evil" in insurance sales had been acknowledged for some time prior, and a solution

had been first proposed by Peters (1894) working with E. A. Wood and the Georgia Life Insurance Company (reported by Gilmer. 1961).

This search for more objective criteria of job performance had been the result of disappointing studies using rating scales as criteria. In fact in the same issue of the Journal of Personnel Research (1925-26) in which the cited studies appeared. an article by Kingsbury was opposing the abandonment of rating scales as criteria. Kingsbury's article suggested that clarifying the concepts of raters, rater training programs, further improvements of scales and consideration of the practicability of rating scales would solve the problems connected with their use.

Hull (1928) devoted an entire chapter (12) to a discussion of the importance and some concepts of criteria of job performance. With regard to the former he says, "...to proceed on a scientific aptitude project without an adequate criterion is hopeless..." and. goes on to present a categorization of criteria as product action and subjective impression. This attempt to conceptualize and systematize job performance measurement was in contrast to naming possible criteria that had been the practice. However even this work tacitly supports the usage of a single job performance measure. ratings. as adequate for criterion purposes. Shortly after Hull's book appeared. Bird (1931) published what is probably the earliest study combining more than two criterion measures and called it an "efficiency index." This index consisted of salary. number of months employed salary increase. number of promotions and ratings by superiors. Today. the hazards of such a composite are obvious but for the time

it represented a departure from the use of a single index of job performance.

The combination of single act or behavior incidents (to receive much attention later) and estimation of an individual average or summarized impressions ignores the scale unit and dimensionality considerations. Early research capitalized on the occurrence of incidents or single acts thus avoiding problems inherent in measurement as well as the abstract problems of definitions. This particular problem, an artificial two category system for classifying, remains today.

Historically, the emphasis was placed on easily identified, specific behaviors or global measures accepted as the composite measures of goodness. With only minor exceptions, the practice continues today in the attempts to predict turnover, lost-time accidents, patent disclosures plus innumerable other points on the continuum. Little or no effort, then or now, is devoted to the identification of the basic dimensions.

Looking back on the period, it seems most peculiar that psychologists did not face the problem of multi-dimensional criteria sooner because it was apparent that others had. Various mathematical models were appearing a short time later that must have been in the germination stages during the period discussed. For example, in 1936, Edgerton and Kolbe, Horst and Hotelling all published studies dealing with combining various criterion measures into a single measure of performance. Travers (1939) described the discriminant function and Wherry (1940), an adaptation of the Edgerton-Kolbe method. All of these studies

had in common the concept that prediction of performance would require a battery of predictors and description of performance would require a battery of measures.

It was during this period that Viteles (1936) introduced a criterion dimension that had received virtually no attention up to the time and has received comparatively little since. It was the satisfaction an individual receives from his work in contrast to the strictly "economic efficiency" aspects of job performance. The issue this raised has continued ever since and only recently has received some consideration as a criterion measure. Even the recent conceptualizations by Herzberg, Mausner, and Synderman (1959) and Brayfield and Crockett (1955) fail to agree as to the relation of attitudes and satisfaction of the individual worker to any operationally defined goals or objectives. To culminate this period, Bellows (1941) published a study that attempted to systematize the development of job performance criteria and Horst (1941) edited what can be regarded as a classic in the field. This latter study, with many eminent contributors and consultants, was a compendium of the problems and techniques of prediction. Written with an eye toward the coming of World War II and its serious manpower problems the study discussed the major problems of prediction of performance and presented the methods for solution as they were known. The study in fact delineates the basic problems of criteria development and performance prediction, many of which are still problems. Emphasized are the complexity of human activities the difficulty of defining

success and that conditions extraneous to the individual can alter his performance. Consideration of these broad areas, with their associated sub-areas implied that extremely complex criteria would be necessary to measure virtually any activity with the needed degree of adequacy. Bellows (1941) op. cit., also delineated some standards by which criteria were to be evaluated, the more important of which were reliability, correlation with other criteria and predictors and acceptability to the job analyst. Nagle (1953) describes the derivation of a composite which was rejected by Guion (1961) as a practical consideration.

World War II brought with it unprecedented opportunities in the general areas of personnel research. Much of this work is summarized by Stuit (1947), Flanagan (1948), and Stouffer et. al. (1949). Criterion development received considerable attention during the course of this war but, under compulsion of immediate necessity, single criteria were commonly used. For example, the pilot and navigator criteria were "check ride" ratings and, for bombardiers, "circular error." These measures had general reliabilities of about .50, .02, and .18. In the case of pilots, it is to be noted that the limit of predictive efficiency had about been reached as shown by Flanagan's (1946) classic study. In this experiment, 1143 persons were sent through pilot training regardless of selection test scores. The multiple correlation for this group with the pass-fail pilot criterion was .66 which, with a criterion reliability of about .50, is very near the maximum possible correlation.

It is regrettable that more attention was not given to criterion development at the time, particularly in view of the fact that some of the more important concepts in need of evaluation had been described by Toops (1944). The article indicated the need for "success profiles" as criteria primarily because success in an activity is not unitary and. further. persons can be successful performers in a given activity for different reasons and at different times. Otis had earlier described this same problem in a book edited by Stead, Shartle, et al. (1949). The detailed resolution was not presented until much later by Toops (1959). However, military studies generally continued to use a single performance measure as a criterion.

The World War II experience did result in a clearer conception of and some work in the general area of criterion development. Stuit and Wilson (1946) published a study showing the marked "influence of the criterion upon the relationship between predictive indices and measure of success." The general point of the study, that continuing attention to better performance measures results in better predictions of performance, is amply demonstrated by the results. In a series of studies, Flanagan (1949, 1954, 1956) had described the conception and refinement of the "critical incident technique" as a method of criterion development as contrasted to criterion selection. In the history of personnel research, this was the first presentation of a systematic method specifically aimed at isolating the bases of performance and. from these, working back toward selection methods. In addition to the critical incident technique,

wartime experience did bring a much clearer recognition and formulation of the nature and characteristics of performance criteria. Thorndike (1949) presented a comprehensive discussion of performance measures. He discussed criteria as immediate intermediate and ultimate. criterion relevance. various types of criteria with their limitations and considerations for evaluating criteria. The study covered most or the facets of criterion development that were and are of importance. Van Dusen (1947) and Jenkins (1946). in a more limited way, covered some of the same material based upon military experience. These studies in criterion development culminated with Nagle (1953) op. cit.. Wherry (1957). and Weitz (1961). The former brings out again the point that individual job satisfaction has had virtually no study as a possible performance criterion and recognizes how introduction of this variable into criterion measures would further complicate predictive studies. Wherry's study stresses the lack of systematic attack on criterion development ahd he says, "If we are measuring the wrong thing, it will not help to measure it better," making the general point of past emphasis on predictors rather than what is to be predicted. Weitz (1961) op. cit.. presented experimental evidence to show how selection of different criteria (in learning word associations) materially changes the interpretation of results and, it is pointed out. that the "laws of criteria" remain to be discovered. Adkins (1947) during this same period discussed some of the assumptions that are made about criteria in predictive studies. One important point was that unless provision

is made for control, motivation, risk, experience personal history items, work environment and other such possible variables are assumed to be equal. On this point, the social scientist needs to refer to Campbell's (1957) discussion of experimental design relevant to variables which affect the outcomes of research. To take one of the variables, motivation, Eysenck (1963) published an experimental study showing that unequal motivation can be extremely important performance variable, and further, it has a nonlinear relationship with performance. It is rare to see a study where the variables named by Adkins are controlled, although they almost certainly have some effect on predictor-criterion relationships.

More recently two other methods, by Lawshe and Steinberg (1955) and Primoff (1957), have approached the evaluation of job performance by first having competent observers rate elements of a particular job for importance or "criticalness." Appropriate predictors are then selected and their relation to the elements determined. After first determinations, refinements are continued to approach the highest possible validity coefficient. This is in contrast to the previously mentioned "critical incident technique" where the approach is to have competent observers report behavioral incidents and, from these, critical requirements are constructed which are to be predicted.

With all this work, has prediction of job performance become any more efficient than it was in the earlier studies cited? A series of studies by Ghiselli and Brown (1951),

Ghiselli and Barthol (1953) Ghiselli (1956) and Balma et. al (1959) indicates that prediction while much more sophistocated has shown little noticeable improvement. The first a survey of studies regarding trainability showed that various aptitude tests tended to be predictive of all occupations at the same level with intercorrelations estimated at .55. The second a survey of the predictive utility of personality inventories showed a range of average correlations of .14 to .36 for eight different categories of occupations. The latter two articles provide some general discussion of the problems that have been encountered in criterion development for years. Such problems as the shortcomings of the various proposed mathematical models lack of functional job descriptions the search for a composite criterion the dynamic nature of jobs the relation of prior experience to the current job and the existence and importance of both individual and situational moderator variables and how jobs differ in different establishments are the more important mentioned. However here and elsewhere there has been it appears a failure to recognize or properly take into account four fundamental problems in the evaluation of performance criteria. These are:

(1) Is job performance reliable? The assumption of reliability is implicit in all predictive studies and must be true if adequate predictions are to be made.

(2) Is observation of job performance reliable? Since all evaluations of performance ultimately rest upon observa-

tion of one sort or another, the question of reliability of such observation becomes crucial to prediction.

(3) Is job performance uni-dimensional? Many studies use a single measurement of job performance (usually a continuum) to evaluate the predicted performance; it is critical to know whether or not such practice can be defended.

(4) Is job performance variability an individual phenomenon? Almost universally individual abilities, traits and characteristics are measured and these are related to some measure of job performance; if there are contingency sources of variance in job performance, they must be measured or controlled for meaningful prediction of performance.

Obviously the above questions have all received some consideration in various research studies. However, it is hoped that a selective survey of the literature will illustrate their overall neglect and, at the same time, their importance. In essence it seems a better understanding of job performance per se will lead to better performance measurement.

The broader problem introduced by Otis (1940), et. al., and Bellows (1941) op. cit., and added to by Nagle (1953) op. cit., Guion (1961) op. cit., and Dunnette (1963;, 1963b), and Weitz (1961) op. cit. is that of criteria for criteria. Certainly practical matters of prediction are of concern, but ultimately some resolution of the abstract problem of definitions and principles must be made.

Is Job Performance Reliable?

Since job performance reliability is fundamental to personnel research, it is disconcerting to find that so few studies have been conducted with the specific aim of determining performance reliability. In addition, many of these have been aimed at determining the reliability of limited aspects or single tasks of a particular job. The task is extremely difficult when the results are intangible or when there is a delay of impact of job performance.

Individual performance variability received some early laboratory attention. Seashore (1931) administered eight motor tests to 50 subjects and, for three, five-minute cycles, 48 hours apart, the reliabilities ranged from .75 to .94. It is probable that these results were inflated by learning, but they illustrate the fact that individual performances vary in reliability. Anastasi (1934) selected 250 Es from an original group of 1000 who were below the first quartile on four tests of a verbal-symbolic nature. The correlations of initial and final scores ranged from .30 to .61 and one of the main findings of the study was that individual variability increased as the trials continued even though individuals maintained their same relative positions. Hertzman (1939) matched two groups of 40 each for general level of ability on the Thurstone Substitution Test but selected one group for high variability and the other for low over the entire test. The two groups varied widely from each other with respect to within-group correla-

tions on subsequent trials with the correlations of the low variability group far more homogeneous than the high variability group. Another interesting point was that as the trials continued the intercorrelations in both groups showed a steady decline. Taylor Hunson and Stone (1945) likewise show an orderly decrement in test intercorrelation as a function of the separation interval. In this study 12 forms of a 250-item number-checking test were administered at 5-minute intervals. The average correlation for succeeding pairs was .925 and declined to .583 with 10 interpolated tests. Cureton (1939), using a longer time interval (5 days), obtained similar results. Owens (1942) gave a group of 15 subjects eight repetitions of seven motor tests. One of the main findings of the study was that intra-individual differences were greater than inter-individual differences. Despite these laboratory indications that even relatively simple task performance was not reliable, the application to determining job performance reliability has been limited; however some studies have been done on task and job performance.

Craig (1924025) reported one of the earliest studies attempting to determine job performance reliability. With "retail saleswomen" it was determined that a "value of sales" criterion had a reliability of .79. Hayes (1932-33) in four studies reports reliabilities of .78, .81, and .87 on first four weeks output vs. second four weeks for various female shop workers and .81 for average "bogey" percentages first two vs. second two weeks all of which are probably inflated

due to the effect of learning. Bellows (1940) reported two studies on operators of card punch machines and coding clerks. With a criterion of errorless production the former showed reliabilities of .89 to .96 and the latter .87. Ayers (1942) used four criteria to evaluate testile inspectors. The criteria with reliabilities by first vs. second week were failure to discover defective units (.73) average hourly production (.85) incidence of units which should not have been put aside (.83) and total units set aside for foreman's decision (.91). Hay (1943) used the control of requiring at least eight months on-the-job before obtaining reliability measures for a group of bookkeepers. On three occasions he correlated first and third days' production with second and fourth with coefficients of .93 .85 and .98. The correlations between the three "occasions" were .83 .79. and .72. Strong (1934-35 1943) in studies with life insurange agents showed that year-to-year production varied with reliabilities of .74 to .84 at various levels of production and another criterion average production of 1926-27 vs. 1929-30 was .81. MacKinney and Wolins (1960) on a year vs. year basis found reliabilities of .45. .25. .55 and .47 for respectively. suggestions submitted by foremen. suggestions installed. suggestions submitted by foremen's subordinates, and subordinates' suggestions installed.

Training research literature provides further insight into the nature of performance reliability in terms of individual dynamics. Smith and Gold (1956) examine the relation of early training performance to post-training performance.

Their results indicate a progressive increase in the correlations between various stages during training and post-training production. They report a range of from .46 between the third and fourth of a 20.5 week program with post-training production to about .82 between the ninth and tenth week of the 20.5 week program and post-training production. A similar effect is demonstrated by the Kornhauser (1923) op. cit. study. Manning and DuBois (1958) employed a unique design to eliminate the effect of pre-training proficiency by using the pre-training proficiency/post-training proficiency regression to obtain a measure of relative gain (residual their term) and found the split-half reliability of total (crude) gain = .56. relative (residual) gain = .57. and final status = .77. Relative gain was considerably more predictable than gross (crude) gain but not as predictable as final status. Fleishman and Fruchter (1960) conclude that early performance in learning Morse code is due to specific aptitudes and later performance probably due to non-aptitude factors such as specific habits acquired during training. Bass (1962) likewise concludes that the decline in test validity over time is due to decreased importance of aptitudes and increased importance of esteem and popularity in sales work. Obviously several factors contribute to the variability of reliability. The impact of the ongoing process on the characteristics of the individual, and the dynamic nature of performance requirements are the two which seem most evident. The problem of temporal proximity well known in educational research, only magnifies the problem of intra-individual variability.

A series of studies by Rothe (1946a. b. 1947. 1951) and Rothe and Nye (1958 1959) was specifically aimed at determining the reliability of job performance in several different occupations. In general this series of studies found individual output to be highly erratic. specific to the individual enormous ranges were found and to quote from the 1958 study. "In this entire series of studies of industrial output the most striking single result is the lack of consistency from time to time especially when there is no financial system in operation. A second important result is the wide range of 'consistency coefficients' of output data. such that a researcher could be entirely misled by tests of statistical significance if he just happened to select a period of unusually high or low consistency."

The findings of Rothe and Nye are supported by others aimed at assessing job reliability. For example. Cohen and Strauss (1946) in an extremely detailed study of performance in a relatively simple task. show that different persons cannot do a given task in the same way. They also found a 1/3 ratio of time. with different methods of doing the same job and say. "From the point-of-view of the methods analyst, there are as many different methods of performance as there are operators." The study casts doubt upon the feasibility of group reliability indices and raises the possibility that the entire question of individual job performance reliability should be re-cast in a unique theoretical context. Perhaps adequate investigation will require longitudinal study of individual subjects. This

approach would control for the interaction of unique individual characteristics with situation characteristics. It is entirely possible for a relatively routine task to vary over time in terms of the responses required. Certainly this is obvious for complex tasks. Carter and Dudek (1947) in a carefully controlled study of navigator proficiency found high reliabilities for single missions but low between missions. in fact, they concluded, "...in many complex skills reliability for any particular trial may be high and yet the correlations between trials, which correspond to test-retest reliability, may be low." That such may be true of other than complex skills is indicated in a study by Klemmer and Lockhead (1962). In the study, of over 1000 operators of key punch and bank proof machines, it was found that individual variability was about 6-10% of the group mean and further that operator variability is relatively independent of mean production level.

A facet that contributes to performance reliability but which has received relatively little attention is that different persons do the same job in different ways. As long ago as 1939, Seashore discussed this aspect. He pointed out that motor, auditory and visual tests show low intercorrelations and personality inventories indicate many possible approaches to problem situations. Walker, et. ak. (1946) tested five experienced pilots for accuracy on 10 different criteria for landing aircraft. Two procedures were used, "Tricks Allowed" and "No Tricks," meaning an individual vs. a standardized landing procedure. The performance of individual pilots showed

more variability under the standardized condition than the unstandardized and under "Tricks Allowed" accuracy of landing was significantly increased. While the scope of this experiment was quite limited it is indicative that job performance among experienced personnel does vary and, in fact, such variability might be desirable. It illustrates once again the point that measures of reliability would be quite different depending upon which aspect of the job happened to be measured.

It is unfortunate that studies of job performance reliability largely must be culled from the literature. However one group of workers, in department stores, has been covered in separate studies that are of interest. Craig (1924-25) op. cit., in a study of 109 saleswomen found a reliability of .79 for value of sales over a period of several months and Stead (1937) coefficients of .83 to .98 over eight objective measures of performance. Otis, et. al., (1940) found, for six measures of job performance, gross sales per day .88 ratio: salary to net sales .83 net sales per day .87 number of sales per day .89 returns per day .75 and actual quota per day .83. The latter study also shows the following table of intercorrelations

	Returns	Number of Sales	Quota per Day
Gross Sales	.58	.47	.65
Returns		.01	.32
Number of Sales			.24

With the high reliabilities found for the variables in the above table and their varying intercorrelations there are obvious implications for job performance reliability. Some of the implications are: how broadly is "job performance" defined how and over what period of time is reliability measured and, possibly, is performance variability an individual characteristic?

The "how" of reliability measurement is directly related to the individual characteristic of variability. The common method for estimating performance reliability is, of course, to correlate two measurements of performance level at different periods. However, the previously mentioned studies by Klemmer and Lockhead, Rothe, and Rothe and Nye all indicated that individual variability is to a large degree independent of level of performance. Coombs (1948) discussed possible different measurements of the same performance but the implications of his study have remained relatively unexplored. Kellner (1960) has shown that the use of "discrepancy scores" in both predictors and criteria results in better performance prediction and has outlined a solid theoretical base for the practice.

Ghiselli (1956) op. cit. in a general discussion of the area virtually dismisses the idea of an index of job performance let alone its reliability and Ghiselli (1960a, 1960b, 1963) has shown that some of the classic concepts of psychometric theory can be seriously questioned when related to job performance measurement. In the latter study it is shown that the classic error of measurement may be better understood as related to traits of particular individuals rather than as a

group concept. The general concept of moderators had been studied by others as Fiske (1957a 1957b) and Berdie (1961). but Ghiselli showed how they could affect prediction of performance. However as applied to performance per se there is little evidence to show the effects, if any. Actually the study of individual performance variability is just beginning, although the problem was thoroughly discussed in a summary article by Fiske and Rice (1955). In their evaluation of the evidence for intra-individual response variability, the authors distinguished three types of response variability. They were, "spontaneous" as might be found with instrumental acts, "systematic" where a response is affected by the preceding response or stimulus, and "variability due to changes" in the subject or situation. One of the major conclusions of the article is that there is a real lack of knowledge in the area, particularly in that of well learned activities.

If we extend our concept of performance behaviors to include acts or incidents which are not directly related to the job functions performed by the individual, we find some interesting but conflicting results. Behaviors such as tardiness, absenteeism, accidents, grievances, supervisory reprimands, and dispensary visits are considered by some to be indications of organization performance (Merrihue and Katzell, 1955) and individual performance (Herzberg, et. al., 1959). Apart from any relation to mental health, the fact remains that each variable is subject to objective measurement. Yet reliabilities vary widely depending upon the situation and the population.

Tardiness, absenteeism, grievances and reprimands seem to be the least stable except over long time periods. On the other hand, accidents and dispensary visits seem to be quite stable with high reliability reported--until the "objective" record is purged of such things as situational hazards, failure to report, inadequate records, etc.

However, with the purified criterion behavior another problem is encountered, in the case of accidents, a shrinking population of "performers." If the cut off point is established as being a chargeable lost time accident, data collected over as long a period as two years still leave, in most cases, the majority of the population in the zero frequency category. The assumptions that the extended time period will provide the opportunity to "act" and that basing research on groups will ferret out the relationships simply beg the question. The fact is that the assumptions are an admission of ignorance or inability to define or measure the performance behaviors being investigated. Psychologists have long accepted either the "J" curve or Poisson distribution as correct to represent low probability single "acts" or incident performances. While this concept does have substantial mathematical support, it seems too parsimonious when applied to situations in which the individual participates purposefully. The restriction in range has its obvious consequences. Extending the time period has other equally undesirable consequences as may the occurrence of the accident itself as postulated by Mintz (1954).

A similar phenomenon is encountered using patent applications as a measure of creativity, or publications, even when

both are corrected for opportunity bias. Taylor (1959) reports factorial reliability estimates (communalities as lower bound estimates) of .23 to .75 for objective indices of scientific productivity and creativity.

It would seem from the foregoing that either measurement is faulty or that measurement is not entirely relevant. It requires only minor immersion in a performance situation to become aware that, with individuals who are not tardy, the time of arrival to work varies considerably, or that among those who do not have lost time accidents there is considerable variation in frequency of cuts, scratches and bruises which do not receive medical attention and are not recorded. Likewise, the scientists who hold no patents may on close examination vary considerably in the frequency of "near" patentable ideas. It seems that major flaws, insofar as reliability is concerned, are in definitions and record-keeping of reasonably important kinds of individual incidents. The data exist; individuals are performing in spite of the failure to measure adequately.

An answer to the question heading this section would appear to be impossible with present knowledge. Actually, as later discussed, job performance is a complex of more or less unrelated tasks, few of which have been measured adequately in terms of their reliability. The correlation of group absolute performance levels affording the classic estimate of reliability actually avoids or at least beclouds the real issue of individual variability of performance. The limited number of studies indicates that individual performance variability is

as much a characteristic of the individual as is an aptitude, personality trait or other more commonly measured characteristic. Actually little knowledge is available as to the extent or importance of individual variability, in fact, it is almost possible to turn the cliche, "more research is needed" into a more pointed "some research is needed," probably using intra-class correlation from analysis of variance designs. If in no other way, it will at least define performance reliability and, it is possible, that individual variability itself may be a better predictor or criterion than those that have been employed in the past.

Is Observation of Job Performance Reliable?

In this section are reported selected studies where the same job performance is evaluated by different methods or by independent raters. This latter point is often difficult to judge from the research report. The authors have probably erred in being overly conservative in selecting studies, but the effort was made to be as certain as possible that the different estimates of the same performance were independent.

An early study by Braunhausen (1929) correlated supervisory ratings with job sample test scores. For two different groups, "Yule's coefficient of association," were 41 and 56. Fay and Middleton (1942), in an ingenious attempt at performance evaluation, obtained recordings of two sales scripts readings by 29 retail salespersons. Each reading was rated, by 139 college students, for (1) enthusiasm, (2) convincingness and (3) sales ability. The following correlations were obtained between ratings of the first and second scripts:

Trait	Males	Females
1	.53	.68
2	.64	.60
3	.80	.71

These studies illustrate a point that continually recurs in the literature; that is, ratings tend to show higher correlations with each other than do more objective measures, and rating tend to fall somewhere between the two.

Comrey (1949) analyzed achievement by West Point Cadets with grade in seven different courses and a composite of rating by peer, academic and military instructors as criteria. A factor analysis of the criteria resulted in eight factors with variance from ratings appearing in only two of the factors to indicate again the relative independence of different performance measures. Ryans and Frederiksen (1951) in discussing the general point of observer reliability cite a study without further identification where raters judging "metal objects" constructed to specifications showed reliabilities of .11 to .55 in their judgments. Use of taper gages in the judging raised the coefficients to .93 and .94. They go on to say, "It is possible to study reliability of performance (as distinguished from judging performance) only where the reliability of judging performance has been shown to be adequate" Gaylord, et. al. (1951), in a study directly concerned with the relationship of performance ratings to measures of actual production found coefficients of .55, .48 and .49 between the former and three indices of production among file clerks. In addition, the raters had production records available, leading to some contamination and probable inflation of the coefficients found.

Peters and Campbell (1955) intercorrelated self and supervisor ratings of proficiency and scores on a diagnostic proficiency test of Air Force mechanics' job knowledge. Correlations ranged from .32 between the second level supervisor ratings and the test, to .37 between the self rating after taking the test and the test scores. Pre-test ratings and first level supervisor ratings were .33 and .35 respectively with

the test. The authors conclude that ratings are not closely enough correlated with diagnostic proficiency test scores to warrant a substitution. To sum up this point, Gaylord, et. al. (1951) op. cit. conclude that the correlation between two criteria should greatly exceed the level usually obtained in validation studies between predictor and a criterion. Their results show correlations of .48 to .55 between composite production records and ratings and .24 to .46 between job elements and ratings.

Springer (1953) compared ratings made by supervisory personnel and by co-workers for promotion to leadman jobs. With a graphic, five item scale, ratings were obtained by 100 workers and, with a graphic, eight item scale, by 68 supervisors. The co-worker reliabilities ranged from .34 to .48, the supervisors .56 to .71 and co-workers vs. supervisors from .15 to .39. In this situation one might be faced with a possible choice between using the one set of ratings or the other. The higher reliability of the supervisory ratings might indicate the choice but Hollander (1954), in various Navy studies, has indicated that "buddy ratings" have been found better predictors for some aspects of performance than supervisory ratings. Hollander (1956), Hollander and Webb (1955), and Wherry and Fryer (1949) rule out postulated contaminating effect of friendship in peer nominations and in fact the evidence suggests friendship may be beneficial, perhaps in terms of opportunity to observe. It is possible that more investigation of this area would indicate that each type of rating would have its place. It

is apparent, in any case, that performance ratings by raters with different points of view have little in common. Liske, Ort, and Ford (1962) found higher interrater agreement of medical student clerkship performance when rater and ratee were in the same specialty. There were no essential differences in faculty rating faculty vs. students rating students. Interestingly though, while ratings were consistent from to time for a composite (Intra-class correlation r_{kk}) interrater agreement (r_{CC}) was only .05 for faculty and .31 for students. The low reliability, the authors conclude, is a function of combining raters and ratees with different specialties.

Some indirect evidence of differences in the perception of the importance of job acts is provided by Prien (1962) and Prien and Powell (1961). In the former, factory foremen and their immediate superiors completed a checklist describing the foreman's job. The average correlation of the relevant pairs (foremen and superiors) was .40. In the latter, training directors and their immediate superiors followed the same procedure and the averaged correlation was .53. Here persons directly involved in the job cannot agree as to the relative importance of duties and virtually must disagree in any performance observations.

Over all it seems evident that the rater must be knowledgeable to contribute real variance in ratings. This general point has received further confirmation in a study by Hicks and Stone (1962). Correlated ratings by peers and supervisors on management personnel showed for over all performance (.51),

promotability (.59) and versatility (.59). While these values are higher than those reported above, they still indicate a real lack of agreement between raters.

Finally, a study by Whitla and Rittell (1953) had 100 mechanics rated on three areas--how well they could get along with others, how well they knew their job, and how well they could do their job--by an immediate superior non-commissioned officer, a flight chief, and first level commissioned officer. Validity coefficients, against a job knowledge test criterion, were .25 to .42 for the first group, .18 to .21 for the second and .20 to .25 for the third group of raters. This includes the correlation for irrelevant measures (getting along vs. test score) which certainly do not appear to differ from the relevant correlations. Similarly Prien and Liske (1962) found averaged correlations over eight graphic scales to be .60 between first and second level supervisors, to .25 between self ratings and first level supervisors and .13 between self ratings and second level supervisors.

Siegel (1954) directly attacked the question of the relationship between various observations of the same performance. In a study with Navy craftsmen performing four tasks, aluminum welding, plastic patching, splicing a cracked aircraft channel, and repairing aircraft fabric were evaluated by a "check list" for each and a ranking of end products by chief petty officers. The inter-examiner reliabilities were .91 to .97 and retests .87 to .83. However, the rho values between check lists and chief petty officer's ratings were for welding .41, patching .16,

splicing .26, and fabric repair .33. It is again obvious that where two or more differently made observations of the same performance are available, the relationship between them is usually low. Siegel, et. al. (1960) found in another much more comprehensive study that ratings by Navy craft supervisors on proficiency and training needed by 70 aviation machinist's mates correlated .35 whereas one would expect a higher relationship on the basis that if proficiency is low, there is a need to recommend training. In the previously mentioned study by Peters and Campbell self ratings correlated with first-and second-level supervisors' ratings yielded correlations of .30 and .23 respectively. The supervisors' ratings correlated .47 for a total sample of 154 mechanics. Although the composite self and supervisor rating correlated .46 with the proficiency test the prediction is considerably short of what could be considered equivalent results.

Bayroff, et. al. (1954) in an experimental study designed to evaluate Army experience with ratings. Some of the relevant findings were that rating ability is a predictable individual skill, several ratings are better than one, control groups should be used to evaluate raters, rater reliability can be assessed properly only by using inter-individual agreement as an index, and that reliabilities tend to drop over a series. Related to this is a study by Bockner (1959) who divided raters into four classes on the basis of the extent to which they agreed in rating the same men. His results showed that higher agreement resulted in poorer prediction of performance in submarine school work. Possibly the clue to these discrepancies

lies in two other studies by Haggerty, et. al. (1959) and Mackie and Eigh (1959). The former obtained ratings of West Point graduates as platoon leaders or company commanders in Korean combat. With multiple ratings on officers who had been in service for several years, the reliabilities ranged from .30 to .63; it will be remembered that the Bayroff study found rating reliabilities tend to fall over a series. The latter study was concerned with Navy machinery repairmen who completed job sample performance tests and relation of these results to ratings. The correlations were .32 and .35 with two school ratings (2 years earlier) on suitability for doing job and .42 with predicted suitability as a machinery repairman. It would appear from these studies that whatever it is that ratings rate is changeable over a period of time and has little relation to objective measures of job performance. It may well be that with changes in skill level or with changes in job requirements over a period of time, the personal behaviors required become more complex, less subject to observation and thus less reliably rated. The concept of the dynamic character of criteria (Ghiselli, 1956) op. cit. is equally applicable to performance behavior. This is particularly attractive explanation if the earlier definition of criterion behavior as situationally determined performance is accepted.

Some general studies covering the problems encountered in job performance evaluation have been reported. Severin (1952) summarized some 150 studies where correlations were reported between different measures of job performance for the same people such as supervisory ratings vs. procuction, tests or

some other measure, associate ratings vs. similar measures and training grades vs. production records. The study can be summarized by the quotation, "The median of all correlations in the table was .28 which seems to be further evidence that one cannot properly substitute one measure of job performance for another without first knowing the degree of equivalence." In this connection, a study by Langdon (1932) is of interest. He reported a correlation of .30 between a work sample test and later piece-rate wages, in a sense, the relation between intermediate and ultimate criteria. Ghiselli and Brown (1951) op. cit., reviewed studies covering some 30 years that reported both training and job performance correlations. The correlations between the two different measures ranged from .15 to .22 for three job classifications and all jobs. Fleishman and Fruchter (1960) op. cit., found correlations of .26 to .41 between successive stages of learning Morse code and conclude that selection tests mainly predicted initial success but later success was more a function of specific habits acquired during training. All four of these latter studies emphasize the desirability of differing methods of assessing job performance at differing levels of proficiency and also raise the question of whether or not the more successful trainees make the more successful later performers. Unfortunately, there is little direct evidence on this question. Milton and Dill (1962) found, however, that later salaries do correlate with starting salary. Again this suggests rather complex phenomena.

Perhaps the most definitive study in the study of performance observation reliability is that of Lifson (1953). In

this study trained time-study personnel rated "work pace" as compared to "normal" by five different persons on four different jobs. Each of these were rated twice at a one-month interval. The "workers" were students who had had industrial experience and who "worked," after considerable practice, paced by a metronome. The study revealed that ratings involve considerable error, some raters rate higher, some workers are rated more reliably, some jobs are rated more reliably, raters tend toward a norm, interactions are of importance, and an analysis of variance showed that one-third of the variance came from rater-to-rater differences. A more recent study by Whitlock (1963) demonstrated a close relationship between reported "effective performance specimens" and ratings. However, the raters knew the individuals about whom the performances were reported and which they later rated. The lack of independence may be the basis for the reported relationship of effective behaviors to higher ratings.

Kipnis (1960) and Taft (1955) op. cit., have discussed some of the major difficulties and distortions that are involved in the observation of performance. Although the former refers mainly to ratings and the latter "the ability to judge others" both seriously question the reliability of human judgments of the performance of others. Taft mainly emphasizes distorting traits within the observers as, intelligence is of some importance in judging others, emotional stability is not a linear but has some relationship to ability to judge, self-insight gives better judgment on any particular trait, "social skill" is an important factor. Others are mentioned but these

are sufficient to show that perhaps studies of raters are needed more than continued performance ratings. Kipnis, in contrast, emphasizes factors more or less independent of performance per se. These are grouped under "External Factors," i.e., propinquity in the sheer physical sense, social setting whether cooperative, punitive or whatever, whether or not criticism is encouraged and "Subordinate Behaviors" as whether behavior "helps" the rater, halo by a subordinate doing well what the rater emphasizes, personal stake by the rater in the rating or its use and various other such consideration.

The studies cited indicate that reliability of job performance observation as presently practiced can be seriously questioned. It is usual to find, where one or more independent observations occur, that the correlation between them is low, especially in situations where an "observation" is some relatively objective measure; for example, a job performance test. The history of evaluating job performance shows the importance of separate measures and limite the value of any studies using a single measure of job performance, even as two raters. It would appear that a major aspect of the "criterion problem" is the fact of unwanted variance and, further, that the sources of this variance are virtually unknown.

In addition to the foregoing information, there is another characteristic of job performance that has been only implicit in the above--the multi-dimensional nature of job performance. The next section will present some of the known information on this topic and how it poses basic problems in the evaluation of job performance.

Is Job Performance Uni-Dimensional

The history of personnel research is studded with the development and use of literally hundreds of performance predictors. In the testing area alone, Guilford (1959) has estimated that 50 of possibly over 100 abilities have been described. In contrast, the majority of reported studies using the predictors have had a single global measure of performance. While it would seem that performance in a particular job is much simpler than the total of individual abilities, is it meaningful to reduce performance measurement to a single measure? In addition, while a particular measure of performance may be identified in several seemingly identical jobs, is it not conceivable that the only similarity is the name given the performance behavior?

The likelihood and consequences of job performance complexity were given early recognition by Kingsbury (1933), "Some executives are successful because they are good planners, although not successful directors. Others are splendid at coordinating and directing, but their plans and programs are defective. Few executives are equally competent in both directions. Failure to recognize and provide, in both testing and rating, for this obvious distinction is, I believe, one major reason for the unsatisfactory results of most attempts to study, rate and test executives. Good tests of one kind of executive ability are not good tests of the other kind." Otis (1953), op. cit., cites a similar example of the college professors who may be equally successful, one on the basis of research competence and productivity, and another on the basis of classroom competence.

Another approach to the study of performance is the direct description of the characteristics of successful and unsuccessful performers. Kerry (1949) and Ghiselli and Barthol (1956) differentiate the successful from the unsuccessful manager suggesting a relation between personal characteristics and achievement. Dalton (1951) on the other hand failed to find a formal pattern of characteristics in career achievement. Informal processes did seem to play a part in career achievement including such things as religion, ethnic background, political belief and participation in accepted organizations. These contradictory results lend little to the concept of individual achievement save to indicate that firm bases for investigation are lacking.

Despite early recognition of the probably existence of several dimensions of job performance, it is only in comparatively recent years that the field has received much attention. Flanagan (1949, 1954a, 1954b) op. cit., has discussed the use of his critical incident technique in isolating and defining "job elements." As previously described, this has been the only systematic attempt to define job performance in terms of its complexity and specifics. However, it is dependent upon observation and reporting of performance as is, and as has been discussed, there is a real question as to the reliability of both. In addition, there is a question of what job performance could or should be which is not investigated with this technique, or for that matter, with any other.

Another approach to defining the dimensions of job functions is illustrated by the studies of Jaspen (1949) and Palmer and

McCormick (1961). Both studies are factor analyses of job descriptions and both recognize their limitations in that they are exploratory. The former study shows six meaningful factors in "lower level" jobs and the latter four in a sample of 250 steel mill jobs. Both of these exploratory studies have indicated that even relatively simple jobs have several independent dimensions and the possibility that more would be found with more rigorous investigation. Studies of the job functions of executive positions by Hemphill (1959) and of supervisory positions by Prien (1963) reveal ten and seven dimensions respectively. It would appear safe to assume that independent functions justify the search for independent performance criteria. Studies by Turner (1960), and Peres (1962), Roach (1956), and Grant (1955) further substantiate the judgment of complexity of job performance however described. These more generally oriented studies have indicated that job performance has a complexity that would require coverage by multiple measurements. This general statement is amplified in what follows by noting the complexity of single measures, single jobs and the relationships of performance indices.

Analyses of single measures of job performance have shown that often they are more complex than seems indicated. For example, analyses of ratings have shown that the intercorrelations of trait scales describe more than one dimension in job performance. Ewart, et. al. (1941) factor analyzed a 12-trait scale and found three factors, Bolanovich (1946) in an analysis found six factors. Taylor and Munson (1951), in a carefully

controlled study, present intercorrelations showing for the most part, low to moderate correlations among separate traits.

Hilton and Dill (1962) op. cit., in an analysis of "salary growth," as a criterion, have shown the considerations that must be given to any single measure for use as a criterion, for example, salary growth is independent of years of employment for the first six years but is highly sensitive to first year salaries. Huse and Taylor (1962) using records on absences for two years on total times absent, total days absent, one day absences and absences of three days or longer found intercorrelations of .00 to .88 among various measures, with absence frequency being the single most reliable measure. King (1960) reported a factor analysis of a 20-item questionnaire covering only "attitude toward company." The study, in ten plants and with 735 employees, found three factors in the one attitude. Eckerman (1948) in a well designed study of employees submitting grievances found 13 items of personal or personnel data that discriminated between grievants and nongrievants. Lurie (1942) factor analyzed 12 indices of occupational adjustment and found three factors in the indices.

From these single measure studies of varying aspects of job performance, it appears that even such relatively simple measures are multi-dimensional in both their behavioral and causal aspects and that global measures of such performance are of doubtful utility.

Even though the multi-dimensional nature of job performance received early recognition, investigation of the dimensions

was later in starting and even yet only rudimentary knowledge is available. An early study by Gottsdanker (1943) is illustrative of the general results obtained when several measures of job performance, particularly those of an objective nature, are used. Using as subjects 44 women learning to operate calculators and as criteria 20 minute tests in a work book, the following intercorrelations were obtained:

	II	III	IV
Test I	.60	.45	.13
Test II		.84	.24
			.38

The tests were simple arithmetical calculations of increasing difficulty and yet the interrelationships, on what would seem to be an easily learned and unitary skill, are quite varied.

During World War II, one of the most intensively studied jobs was that of learning to fly aircraft. With the pass-fail criterion, Guilford (1947) showed that eight factors were involved in this single criterion of performance. Further analyses of the same job by Dudek (1949) and Michael (1949) compared factor loadings in the criterion with different populations. The former used two groups of pilot trainees and one group of women trainees, the latter, two groups of white and one Negro group of trainees. Both studies found that the factorial description of the criterion varied from sample to sample. The variability was not only in weights but the appearance or non-

appearance of different factors. An investigation by Fleishman and Ornstein (1960) indicated that such global measures may be even more complex than is shown by these studies. Sixty-five flying student were tested on 24 strictly flying maneuvers. A factor analysis of the maneuver score intercorrelations revealed six factors in the maneuvers. When it is considered that maneuvering is only a limited aspect of the aircraft commander job and such maneuvering is factorially complex, it can be surmised that the composition of the entire job is factorially formidable. As a sidelight to the cited study by Fleishman and Ornstein, it might be noted in reference to performance reliabilities that the reliabilities of individual maneuvers, as estimated by the communalities, varied from .20 to .77 with a median of slightly below .60. In view of the studies cited, two quotations from them are pertinent, i.e., from the last, "Similar analyses of the interrelationships among component performance measures of other complex jobs may provide one way of defining the ability requirements underlying proficiency in those jobs." And, from Michael, "It is quite probable that the gross pass-fail criterion could advantageously be replaced by many independent and relatively pure criteria." Trites (1959) et. al. attempted to do this. In his study, an analysis of performance for general flight training revealed five factors extracted from 22 performance measures of which only one was actual flying.

An area of performance that has received comparatively more attention than others in terms of its dimensionality is

that of academic achievement. Gaier (1952) studied criteria for success in medical school by analyzing grades received by two different classes. The results were "equivocal" because it was determined that the classes were not equal in either ability or achievement and further it was believed that the standards of evaluation varied from class to class. This point is substantiated by Aiken (1963), who presents results indicating that the concept of the average student is a function of the level of performance of the group and is not a stable abstraction. However, Haier (1952) op. cit., indicates that success was based upon ability, motivation and work habits and adequate prediction would require broader criteria to allow all three to function. Studies by Locke (1963) and Prien and Lee (1963), op. cit., of school achievement, indicate at least two dimensions; namely, structured achievement and unstructured achievement. Additionally, Prien and Lee note a social achievement dimension. Preliminary studies by Davis (1964a, 1964b, 1964c) analyzing faculty and student perceptions of performance indicate considerably greater complexity.

Newman, et. al. (1952) studied two classes at the Coast Guard Academy. The criteria consisted of ratings by peers, officers and staff both ashore and at sea, "demerit scores" and course grades yielding over 20 measures for each class. Cluster analyses of over 2,000 correlations revealed three independent clusters of general adaptability to Academy life and activities, physical proficiency and attitudes, and academic grades. Since the results for two separate classes agreed, it was concluded that the results seemed definitive.

Graham (1954) and Bair, et. al. (1956) both analyzed achievement in Navy flying training. The former with seven criteria from both pre-flight and flight training obtained four factors in achievement. It is interesting to note throughout the table that two measures of flying ability have virtually zero correlations with all other measures and agree with each other to the level of .28. The latter study with 12 measures of achievement in pre-flight resulted in only three achievement factors. Of course this study also showed higher intercorrelations since only grades were measured, but even here the highest correlation obtained was .72 and that was final Navigation grade with the summary grade measure and is somewhat spurious.

Another study of achievement in the Coast Guard Academy by Pettner, et. al. (1959) had as criteria ten academic grades and ratings of cadets on cruise. Factor analysis of these, along with 20 tests, showed that criterion scores had significant loadings on six of 15 factors extracted. This analysis, more detailed than that of Newman, reveals six distinct bases for academic achievement and rating in one aspect of a total job.

These studies of academic achievement performance measures again reveal that even rather limited aspects of a "job" are complex. Unfortunately, none of these studies as yet have reported comprehensive follow-ups of later careers and thus, the relation of training achievement to job proficiency is not clearly established; it can only be surmised that this performance would be even more complex than training alone.

Some insight as to the complexity of cumulative job performance data is provided by Richards, et. al. (1965) reporting a study of medical specialists. Eighty performance scores including three measuring academic performance were factor analyzed. The analyses yielded 29 factors, and this is viewed as a conservative estimate of the complexity since no attempt was made to measure patient responses or the quality of medical care. Of particular interest, though, was that poth pre-medical and medical school performance were independent of the job performance of the group studied. The above study is perhaps the most comprehensive one performed to date and clearly illustrates the magnitude of the criterion complexity issue.

Performance measures for one job area have been subjected to several analyses--the sales job. Rush (1953) has presented what can be regarded as a classic study in the field, or in all personnel research for that matter. The investigation covering both preliminary and cross-validation aspects used criteria of percent of assigned quota achieved, average number of sales, average monthly volume (all corrected by a base sales figure), grades in a technical sales school and supervisory ratings on a nine scale form. From a table of intercorrelations, Rush extracted four factors of, I - objective achievement with loadings on the described indices, II - learning aptitude with loadings in grades and ratings of technical knowledge and learning, III - a general reputation (halo) factor with loadings in ratings and IV - a sales technique and achievement factor which had the only communality between objective sales measures and ratings, much weakened, however, because of rather low,

scattered loadings. From a large number of predictors including aptitude tests, a personality inventory and personal history items, multiple regression equations were constructed to predict each factor. Only IV did not produce a significant multiple R, and the best predictors for each were different in every case. It is of interest to note that on the predictor, "number of accounting courses," which had been used as positive actually had a substantial negative relationship with Factor I in the later analysis. This study illustrating the multidimensionality of job performance, as is specifically commented in the study, also embodies some other points previously mentioned, as the relatively low relationships of varied performance measures and the lack of relationship between objective sales measures and ratings of sales ability. It would appear there are actual achievements in both sales and training and then an unrelated supervisory opinion concerning the achievements and, the suspicion exists, that this is common to many other fields of work.

Two other studies with objective measures of selling performance by Kirchner (1960) and Miner (1962) present tables of intercorrelations of various performance measures showing relatively high, positive correlations among them. These would seem to indicate the possibility of a single "selling ability" factor. however, another study by Baier and Dugan (1957) using 13 objective measures of sales achievement by insurance agents presents a table of intercorrelations that obviously contains more than one factor and indicates that selling ability in at least one field is not the unitary ability that might be supposed but is more accurately described by Ruch's study, op. cit.

Of course, sales achievement has been so little explored, in the sense of these later comprehehsive studies, that only the most tentative judgments are possible; however, it does appear that some general principles have emerged. As indicated in an early study by Dorcus (1940) the establishment of an "objective" criterion is a quite ticklish procedure and overlooking even seemingly minor aspects can apparently seriously bias such criteria. In fact, Dorcus constructed "economic maps" of sales territories to furnish base points. Related to this point is the sheer number of criteria; if relatively few are used it appears there is a greater tendency for them to be more closely related in a positive manner, perhaps the result of a limited view of the actual possibilities. Finally, the temporal aspects of the stability of relationships are virtually unexplored.

One type of investigation that can perhaps illustrate the multidimensional nature of job performance better than any other is that where criteria of quite diverse nature are specifically investigated or where values of equally diverse performance predictors are assessed. Gadel and Kriedt (1952) in a study of 193 IBM operators determined job satisfaction and interest by questionnaires and job performance by supervisory ratings and obtained the following intercorrelations:

	Performance	Satisfaction	Interest
Satisfaction	.09		
Interest	.08	.44	
Aptitude	.41	-.11	-.11

Ferguson (1960) in a study of the utility of aptitude, interest and personal history items ("Economic Maturity") in predicting the job performance of insurance agents concluded that personal history and aptitude items predicted performance whereas, survival was predicted by interest. He also hypothesized that aptitude is a joint function of interest and ability and that long term prediction will depend much more upon interest than upon ability. Clark (1961) in a quite comprehensive study of several Navy technician groups found virtually zero correlation between aptitude and interest measures and yet substantial validities for both in the prediction of technical school grades.

These selected studies indicate that performance is complexely based in the individual himself and, it is to be presumed, results in complex effect upon job performance. It will require established performance measures before the functioning and relative importance of these variables within individuals can be determined with any degree of accuracy and comprehensiveness.

Other studies have demonstrated points that are of interest in this section. Bartelme, et. al. (1951) using a version of a driving skill test, developed by the American Automibile Association, attempted to predict Army truck driver performance. The interesting point is that the test battery predicted the criterion to the extent of .24 for light, .11 for medium, and -.12 for heavy vehicle drivers. If a generalization is warranted on the basis of a single study, it would be indicated that a rubric covering a job as here, truck driver, must be carefully investi-

gated before being accepted. Lawshe and McGinley (1951) in a study of proof reading performance found a correlation of .06 between productivity and errors to indicate the likelihood of these being independent measures of achievement in a single job.

Another approach to determining the dimensions of job performance is that of using what might be called organizational indices to evaluate performance. To illustrate, Clarke (1946) found a correlation of .52 between absenteeism and turnover which obviously is much higher than many attempted predictions of turnover. Palmer and Schroeder (1961) show that theft of company materials is inversely related to the practice of allowing employee discounts. Comprehensive or conclusive studies to identify basic dimensions which could be ascribed to the organization are not available.

Heron (1954) used six criteria to evaluate the performance of bus conductors. They were Gross Earnings, "Shorts" on cash for tickets sold, number of periods of absence, disciplinary actions, times late for duty and a supervisory rating on how much the employee was a "source of concern" to his supervisor. The intercorrelations and a factor analysis revealed:

	2	3	4	5	6	I	II	h^2
1. Supervisory Rating	304	505	382	127	485	70	03	493
2. Gross Earnings		095	414	057	241	44	-42	372
3. Shorts			272	230	447	61	32	473
4. Absence				018	369	56	-33	426
5. Disciplinary Action					274	28	27	148
6. Lates					---	70	14	513

A study of skilled tradesmen by Ronan (1963) factor analyzed 11 job performance variables. Included were ratings as well as personnel file data on organization type indices. The analysis revealed four factors in these various indices of performance. A study by Fleishmen, et. al., (1955) used three of the same variables derived in exactly the same way, i.e., absenteeism, accidents and grievances. The table below shows the comparative correlations as study I, the former, and II, the latter:

		2	3
1. Absenteeism	I	.05	.24
	II	-.20	.37
2. Accidents	I		-.06
	II		-.18
3. Grievances			----

A similar comparison for the common variables for the Heron and Ronan studies shows:

		2	3
1. Supervisory Rating	I	.52	.29
	II	.38	.13
2. Absence	I		.18
	II		.02
3. Disciplinary Action			---

These three studies seem to show considerable stability of relationships over widely varied organizations and populations despite relatively low reliabilities for some. The reliabilities,

estimated from the communalities, in Heron's study were .493 for rating, .426 for absence and .148 for disciplinary action. The same in Ronan's study were respectively .543, .612 and .232. Fleishmen, et. al., obtained reliabilities, corrected by the Spearman-Brown formula, of .85 for absence, .72 for accidents and .73 for grievances. All of these studies covered comparatively long periods of time which may have allowed relationships to appear that ordinarily do not do so. Research by Penn (1955) on the reliability of accidents indicates that reliability increases as the duration of exposure increases and as hazard increases. For the high hazard group maximum reliability was reached during the second of a four-year period. For the medium hazard group increases were found through the third year. The low hazard group first year reliability was not significant ($r = .09$, $N = 60$) but reliability increased throughout the period. Maximum reliabilities for the high, medium, and low hazard groups were .87, .87, and .55 respectively.

It would appear that further studies of these "organizational indices" might well be fruitful in attempted criterion development. Whatever else might be said they do evaluate "real" aspects of performance as contrasted with ratings which may or may not be doing so. It is possible that further development might lead to a partial, ultimate criterion if the contradiction can be accepted. For example, a dollar value could be estimated for absenteeism, accidents, turnover and many other such measures. In this way, for some dimensions of performance, the "dollar criterion" concept of Brogden and Taylor (1950) might be approached. This

probably would require more limited statements about validity, but if all these independent performance indices are to be predicted, the predictor battery likely would be immense. Another problem would be the delayed impact phenomena characteristics of higher level jobs. The true and complete impact of a mode of performance, an act or a sequence might not occur for years. Even when the results become manifest, the question of assigning responsibility would remain.

An indication of the limitations of the studies cited immediately above and the complexity that might be encountered is found in a study by McQuitty, et. al. (1954). Here behavioral descriptions of best-average-poor aircraft mechanics were obtained from peers and supervisors. From these a "descriptive inventory" was constructed and 428 line supervisors rated some hundreds of mechanics. A factor analysis of the ratings extracted 23 factors which accounted for only 50% of the variance. Obviously there are a large number of relatively independent behavior dimensions related to job proficiency. This study found interest, character, personality and aptitude measures of importance with only the limited criteria of supervisory ratings. If, in addition, other criteria were to be used, the task of isolating all the possible relationships appears staggering if indeed it can be done in the foreseeable future.

In summary, the information presented in this section indicates beyond any doubt the multi-dimensionality of job performance; in fact, the phenomenon is characteristic of even limited aspects of a job as was shown in flying an aircraft.

To attempt to evaluate job performance with a single measure is worse than useless, it is misleading; and, for ratings, to keep in perspective all dimensions of performance while rating would appear impossible.

Is Job Performance Modified by Extra-Individual Conditions

Tacit in the design of most research studies has been the assumption that job performance is directly the result of characteristics of the individuals involved. Predictors of many sorts have been used to describe individuals with the results related to some performance criterion but it is rare to find, in any single study, an attempt to determine if intra-individual are the only sources of variability in performance.

The possible existence of biasing conditions within the situation has been called the single most important criterion problem by both Bechtholdt (1951) and Cureton (1951). As Anastasi (1950) has pointed out, even though the shortcomings of any criterion are known, the operational results are, that if it must be used, only the interpretation of validity coefficients would be changed, there still remains the relationship of predictor and criterion behaviors. Kipnis (1960) op. cit., has presented evidence to show that performance ratings are distorted by supervisor-subordinate relationships and the context in which they occur. Katzell (1962) has pointed out the general inadequacies of present day organizational theory in any attempt to assess effects of job performance as a dependent variable. The most comprehensive statement of the considerations in this area is that of Horst (1941) op. cit. In stressing the lack of and need for research on behavioral fields the statement is made, "Without dwelling further on the point, it is clear that an individual's performance in an activity cannot be viewed

as an isolated phenomenon outside the environmental context in which the activity takes place. The activity must be analyzed, not only in terms of the characteristics of the person engaged in it, but in the light of the principal external conditions which may influence it."

The general trend of opinion and what little evidence exists seems to be that situational conditions can modify individual job performance. To result in such behavioral changes, it is necessary to assume that the individual has in some way been changed. It seems unlikely that such conditions could alter individual aptitude, ability or interest (in the sense of interest inventory measurement) levels, the changes must have an attitudinal, motivational or some such taxonomic base. It would seem that such changes would result from reactions or perceptions largely based upon personality traits. One of the assumptions is that attitudes, and specifically job attitudes, are in some way related to personality characteristics. These relations were established in a study by Svetlik (1961) but only with any definite degree for these individuals manifesting some type of career concern (voluntary referrals for vocational counseling).

The whole area of personality traits, attitudes, morale and their relations to criteria of performance has proved extremely difficult to attack and as yet only the most tentative results have been obtained. An early study by Lurie (1942) op. cit., found three factors of occupational adjustment from a factor analysis of 12 indices; however, this general approach has not been directly followed by similar studies but has been approached in different ways.

An experimental study by Heron (1952) used scores from 22 personality tests given to 80 unskilled (pour lead in molds) workers. The criteria were average productivity for 67 weeks and ratings by six supervisors, the correlation matrix was factor analyzed. Four factors were found but none were related to the criteria. What was found is quoted, "It seems in the sample studied it is the heterogeneous group of relatively unstable men who tend to be a source of concern to their supervisors." This is based upon the fact that two of the factors correlated to the extent of .53 with "job adjustment."

Peck and Parsons (1956), using Worthington's application blank as a diagnostic instrument, found relatively high correlations, up to .77 rho, with production and "favorable" personality traits. They also found, as an incidental comment on performance reliability, that high producers showed little variation of production whereas low producers were much more variable. There is some possibility of criterion bias in this study since employees were working against production standards and, in such situations, it is quite common to find agreed upon levels of production. It is also worthy of note here, as in Heron's study, the persons with unfavorable personality patterns where chronic supervisory problems. Heron has suggested the possibility that poor job performance is the independent and poor adjustment the dependent in the work situation. Such an hypothesis is not beyond the realm of possibility, but there is little supporting evidence.

A partial clue to the discrepancy between the two studies cited above may be found in the results of two other studies.

Kipnis (1962) using a specially developed test, found "persistence beyond minimum standards on tiring tasks" did predict performance particularly among lower aptitude individuals. Eysenck (1963), op. cit., has shown that motivation has optimum levels and that its relationship to performance is curvilinear. From these studies, it would appear that performance can be affected by personality traits but, in general, the relationship is quite complex and may be in the nature of moderators. Parenthetically, the existence of "trouble-makers" in industrial organizations has often been doubted, but it appears from the first two studies that they do exist and, whether or not their performance is good or poor, they can be undesirable employees on another dimension of job performance.

Regardless of whether one is willing to attribute performance effects by the variables considered in this section, the fact remains that a great deal of effort has been expended in their investigation. There have been, first, studies of organizational features or characteristics as shown by their relation to various objective indices of performance.

Using one index, turnover, Parkinson (1928) in a study of 99 selling and distributing organizations with over 60,000 employees found higher turnover related to larger organizations. He says, "The outstanding observation from the canvas is that personnel conditions (labor turnover) are least favorable among the large forces as a class, and the most favorable among the small organizations." However, Parkinson specifically points out that such a generalization is not completely warranted since there is considerable overlap and some large organizations

have distinctly favorable turnover rates. Savatsky (1951) presents data to show that in larger departments, within a single organization, the turnover rate is higher than in the small departments. Greyston, et. al. (1952) have presented data to show that the question of turnover as related to organizational size needs further investigation since no distinct linear trends regarding company size or department size were evident for employees of either sex.

These data, along with the qualifications mentioned by Parkinson, indicate that while a relationship between organizational size and turnover does exist, it is by no means a simple one. In any case, using turnover as criterion would require taking into account organizational size as a factor affecting performance, but the contributing effects of other influences would also need to be investigated.

Another single index that has received some study, as affected by extra-individual factors, is that of accidents. Kerr (1950) and Keenan, et. al. (1951) both have studied organizational characteristics as related to accident rate. The former concerned 53 departments in one company with conclusions, stated as only tentative, that accident frequency is greatest in those departments with "lowest intra-company transfer mobility rates, smallest percent of employees who are female and on salary, least promotion probability for typical employee, and highest mean noise level." For severity the findings were, "...heavily male in sex ratio for salary as well as production personnel, low in mean promotion probability, low in fertility of suggestion field, low in employee suggestions contributed, high

(relatively) in average employee age level, and higher in average employee tenure." An incidental finding of this study were correlations between accident frequency and turnover of .03 and with severity -.20. The latter study covered 1,945 lost time accidents, 7,108 employees and the years 1944-48. With supervisors rating "departmental conditions" of 44 departments, the tentative findings were that promotion probability increases safe behavior, "comfortable shop environment" is a major determinant of safe behavior, crew work brought higher injury rates as did greater manual effort involved. As in the case of turnover there is evidence to show that a moderator is at work with both injuries and lost time accidents as a measure of job performance. A possible moderator suggested by Kerber (1958) is that employees on an incentive pay system may have fewer reported injuries as evidenced by dispensary visits than do day-rate workers simply because minor injuries are economically costly to the employee. Kossoris (1940) using data supplied by the Wisconsin Industrial Commission, the Swiss National Accident Insurance Fund and the International Labor Office Study of Austria studied the frequency rate of accidents as related to age over some 500,000 industrial accidents. He found older workers less susceptible to injury than younger workers, older workers had more serious accidents and required longer to recover. As in the case of turnover as a job performance criterion, it would appear that accidents are to some extent related to situational characteristics but, as yet, the detailed relationships are hazy.

An interesting study by Marriott (1949) evaluated work group size as related to output measured by piece work earnings per man. In two separate studies with 153 groups and 79 to 98 groups, low, inverse correlations were found indicating smaller sized groups generally showing larger output. One exception was that groups of over 50 showed larger output than those immediately smaller, presumably because this is the point of mechanization and/or the group has become so large that its influence has decreased. Actually there is little information related to this particular topic, the more recent burst of activity concerning small group effects having largely been confined to decision making, assumption of leadership or similar topics.

The existence of another consideration in studies of the type presented above was shown in a study by Ferguson (1951), op. cit. In comparing validities for the Life Insurance Aptitude Index, a wide variation was found across districts even though score distributions were comparable. It was apparent that the evaluations of job performance were quite different by the managers of different districts or agencies.

Cureton and Katzell (1962) in a study of 72 divisions of a company using five measures of divisional performance and five descriptive situational variables found two factors showing that a non-urban culture pattern reflects small plant and community size relatively higher productivity and profitability whereas the other, urban, shows lower wages, fewer female employees, no union and higher turnover. Thus, one aspect of an orga-

nization, situs, had differential effects on several measures of overall performance.

These studies of specific aspects of performance or specific independent variables indicate that situational variables probably have some effect on job performance; however, there are indications that these single aspects are acting in larger contexts. Recently this aspect of job performance measurement has received increasing attention and more comprehensive studies have been completed.

Stodgill, et. al. (1955) studies the administrative behavior of 470 Navy officers in 45 positions, 47 organizations and from Ensign to Admiral. A factor analysis of the data yielded eight factors that tended to group individuals by the type of position they held. It was also found that types of positions tend to be found either in small or large organizations or in ship as opposed to shore units. It was clear from the study that performance is at least in part determined not only by job demands but by particular job and place. In terms of job performance measurement it was suggested that measures of job performance patterns might be devised as opposed to evaluation of such traits as initiative, judgment, etc.

Turner (1960) op. cit., in a study of foreman performance in two different plants, factor analyzed two matrices of intercorrelations obtained from 11 objective measures and a nine trait rating of job performance. For both plants three similar factors emerged covering rated performance, probably halo and reputation, an employee relations factor and a bi-polar

factor covering scrap and suggestions indicating that good performance on one is accompanied by poor performance on the other. Two other factors were much more poorly defined, and different in the two plants. One might possibly be "structuring" as described by Fleishmen, et. al. (1955), op. cit., but the other seems to indicate specificity to a particular plant. Again was found in this study the lack of relationship between ratings and objective measures as previously discussed, relatively low reliabilities of certain individual measures, particularly if measured over short time spans, and the multidimensional aspects of performance assumed by the author, "It appears there is more than one pattern of foreman success and that it may be unrealistic to expect foremen to do well on all aspects of the job." This study does indicate the possibility of establishing at least some performance indices that are common in various organizations, but it also indicates the possibility of specificity to a single unit.

Thorpe, et. al. (1961) in a comprehensive study of three Air Force career fields have delineated the complexity of approaching job performance from both total performance and organization points of view. In this study, several measuring instruments were specifically developed for the study. They were an opinion inventory to measure job satisfaction, an effectiveness rating scale and a specialized interview with supervisors. In addition, peer nominations were obtained along with a host of personal history items, aptitude and achievement test scores and indicators of military achievement. In-

tercorrelations and factor analysis yielded six performance factors. These were: (A) - General Competence, (B) - Promotion Potential, (C) - Career Orientation, (D) - Peer Recognition, (E) - Job Satisfaction, and (F) - Job Centeredness. A parenthetical point of interest is that aptitude scores were related only to "A," the General Competence factor. From the point of view of this section two quotes by the authors are pertinent, "...seem to indicate the need for multi-dimensional evaluation of airmen performance." and, "There was considerable similarity of loadings across the three career fields to suggest that a universal criterion for job evaluation is possible." Other tentative results of the study were that it may be possible to determine training needs with this procedure, discover supervisory potential early and that only six scores would be needed to predict the six factors.

Seashore, et. al. (1960) presented a study that would seem to cast some doubt on the utility of measuring across organizations or jobs and the use of various organizational indices as criteria. The study used as criteria over all effectiveness (a rating), productivity, chargeable accidents, unexcused absences and errors. the latter four, objective measurements. In the evaluation, three hypotheses were evaluated, (1) intercorrelations of job performance measures will be consistent, (2) patterns of intercorrelations among the variables similar in size and sign as between individuals and organizational levels of analysis and (3) relationships among job performance criteria for individuals in any one organization are repre-

sentative of relationships over a set of homogeneous organizations. The authors found for (1) that three of five criteria were internally consistent and the hypothesis received some support, for (2) the results were inconclusive and for (3) rejection as "one must conclude from this evidence that the relationships among various aspects of job performance are highly variable..." In evaluating the study it should be pointed out that data were collected for a period of only one month. Turner (1960), op. cit., says, "Single monthly scores on criterion measures tend to have inadequate reliability across time. Averages of several monthly scores are needed to attain a satisfactory level of reliability." Turner bases this statement upon his study where reliabilities over one month ranged from .03 to .59 with a median of .35. Over 3 1/2 to 5 months, reliabilities as estimated by communalities presented by Turner, were from .14 to .92 with a median of .82. Most of the higher, of course, were in ratings but even the objective measures had a median in the .60's. Penn (1955) op. cit., indicates the increase in reliability of accidents leveling off after 1 year for high hazard jobs and still increasing at the end of four years for low hazard jobs. This temporal aspect of job performance is one that has received very little attention, longitudinal studies over any time periods exceeding one year are the exception. It might be well to take up the topic here.

Viteles (1929-30) found over almost a two-year period that substation operators, classified into three groups by 13 supervisors, confirmed the classification using an "error" cri-

terion, the poorest group having over seven times as many errors per man as the best group. Here, in contrast to previously cited studies, a rating is confirmed by an objective criterion but also in contrast by records kept over a comparatively lon- period of time.

Another early study by Ball (1938) found a correlation of .71 between mental ability, as measured by a relatively simple test, and occupational status of office workers after an 18-year period. Further, there is no evidence of contamination in the study; it appears that the coefficient found is a good estimate of the relationship. Stead (1937) op. cit., in a study of department store sales personnel used eight objective measures of performance on a year-to-year basis. He found reliabilities of .83 to .98 and a multiple correlation of .65 for six tests, with a combined criterion. Strong (1934-35), op. cit., in a study of insurance sales as a criterion found reliabilities of .77 to .84 on a year vs. year basis and in another study (1943) op. cit., found reliabilities of .74 to .84 correlating two years production (1926-27) with production for the years 1929-30. Knauft (1955) correlated test scores from a general mental ability test (LOMA-I, 15 minutes) with job level obtained over a 17-year period. A correlation of .60 was obtained over seven job classifications and, further, this is uncorrected for a restriction of range which would probably raise it to near the value Ball, op. cit., found for similar period of time.

Whitlock, et. al. (1963) in a study relating "unsafe behaviors" to incidence of accidents specifically studies, as one

facet of the investigation, the influence of time on the relationship. The trend of the data is for the relationship to increase with time. A specific recommendation of the study is that investigations in this general area of job performance must allow sufficient time for relationships to become apparent.

In contrast to these studies, others showing longer time periods, tend to attenuate relationships. The study by Bayroff, et. al. (1954), op. cit., presents evidence to show that reliability of ratings tends to drop, this with four ratings over a period of weeks. Ghiselli and Haire (1960) studied taxi cab drivers over the first 18 weeks of their employment and found, in general, validities dropped, no single consistent predictor and validity correlations change when different criteria are used. Bass (1962), op. cit., found that ratings of sales personnel showed lower relationships over a 42-month period.

Actually these latter studies concerned with ratings and relatively short time periods serve to emphasize the need for longer time periods and the questionable utility of ratings as criteria. The longer studies previously described using more objective criteria show quite substantial relationships even with simple predictors and high reliabilities for the objective performance measures. Studies of the kind are, of course, difficult to conduct because of the necessity of record keeping, the influence of learning with new employees, the ever present danger of contamination, and sample attrition, however it appears, from the limited evidence available, that more

studies will need to be conducted for a full appreciation of job performance and criteria development.

An area where the influence of extraneous factors on job performance has been extensively studied concerns that of leadership or supervision effects on morale, attitude or more directly some measure of job performance. For example, Matthews (1951) after a review of studies of leadership up to that time reached as a partial conclusion, "Intercorrelations among various measurements of leadership were low but positive. There seems to be some tendency for those who are leaders in one situation to be leaders in other types of situations. However, a considerable portion of the leadership variance cannot be attributed to persons but probably must be attributed in part to situations," and "it will be well to recognize that there are probably certain general requirements and also that there are certain requirements which are unique for the particular leadership situation one has in mind." Matthews also points out that up to the time at which he was writing there were few studies to show the effects of leadership on performance, primarily because of lack of suitable criteria.

Fleishman, et. al. (1955), op. cit., in a comprehensive study of industrial foreman leadership isolated two factors called, "Consideration" and "Initiating Structure." Essentially the former factor describes, "a more friendly, trusting person who develops a certain warmth between the leader and the group," while the latter factor describes a person who is more prone to define his relationship to the group, roles he expects to be played and organizes the job. The scores for each

of these were correlated with proficiency (management rating) and four objective indices of performance in both production and non-production departments. The results are shown below:

Consideration	Proficiency	Absenteeism	Accidents	Grievances	Turnover
Production	-.31*	-.49*	-.06	-.07	.13
Non-production	.28	-.38	-.42**	.15	.04
Initiating Structure					
Production	.47*	.27**	.15	.45*	.06
Non-production	-.19	.06	.18	.23	.51**

* - significant at 5%

** - significant at 1%

It might be mentioned that reliability correlations, as measured by separate administration, for the leadership designations were .58 for Consideration and .46 for Initiating Structure. The study also found that workers liked a foreman high on Consideration but a foreman is considered more proficient by superiors if he is higher on Initiating Structure, "consequently there appears to be a conflict between morale and efficiency." We have here again a situation where by performance, i.e., Consideration, a foreman might reduce absenteeism, accidents, grievances and turnover but in the opinion of his superiors he would not be proficient on the job. This study also points out the situational variables in leadership, different leaders may be required in production and non-production departments. One point the study most forcefully indicated was the extreme complexity of the leadership-job performance relationship.

Clevin and Fielder (1956) using an instrument to measure "ASo" score, i.e., supervisory prediction of subordinate's behavior, which dichotomized supervisors into more accepting, approchable individuals as opposed to more critical, analytic persons found proficiency of work crews under the latter to be much more predictable. This was true of supervisors in more direct contact with the crews while supervisors more distant from the crews, or work site, did not show such predictions. The study is of particular interest because it covers a longer time period than usually found and, the criterion, tap-to-tap time of open hearth heats is almost completely objective. In addition, the odd-even months criterion reliability was found to be .82. The finding by Clevel and Fiedler agrees with that of Fleishman, et. al. (1955), op. cit., in that the supervisor showing Initiating Structure is regarded as more proficient by management. However, this work group also shows higher rates on some undesirable indices, for instance, accidents. It appears from these studies that method or techniques of supervision have some influence on job performance, but they have their effects in complex, in fact, contradictory ways. This is further supported by the studies of Turner (1960), op. cit., where bi-polar factors were found in foreman performance.

Two reviews of the literature of this area, Brayfield and Crockett (1955), op. cit., and Herzberg, et. al. (1957), have arrived at somewhat different conclusions. The latter cites 26 studies covering the relationship of satisfaction or attitude to productivity. It was found that 14 showed a positive rela-

tionship, nine showed no relation and three a negative relationship and, it is concluded, that supervision "definitely affects" productivity to some degree. The former concluded, in general, that any relationships were quite nebulous, in fact, the efforts in the entire area were seriously questioned on such bases as sampling involved, inadequate criteria, bias of self-report and group statistics. On a theoretical basis, it was also questioned why morale, attitude, etc., should be related to productivity, no one-to-one relationship has ever been clearly established. Further, the complexity of human goals, needs, satisfaction and such designations when placed in a complex situation of a work system have been most inadequately explored. Such an analysis would involve individuals, the factory social system, the work group, union and community at large. The authors said, "We seem to have arrived at the position where the social scientist in the industrial setting must concern himself with a full-scale analysis of that situation." and, "Pursuit of this goal should provide us with considerable intrinsic job satisfaction."

This complexity, not intrinsic job satisfaction, was indicated in a paper by Kahn and Morse (1951) which indicated the probable dimensions of individual satisfaction, the independent variables, the uniqueness of individual needs and the likelihood of interactions in a work situation. To some extent attempts at systematic investigation of these have been made by the Survey Research Center at the University of Michigan. In summarizing some of the studies, Maccoby (1949) tentatively

concluded more pressure from above exerted on a supervisor gave lower productivity and, supervisors assuming a "leadership role," gave higher productivity. Mann and Dent (1954), reporting studies by the same group on what makes an effective supervisor, found "employee orientation" important by both subordinates and management (contrast this with the previously described Fleishmen, et. al., study) and, with Likert (1951), the importance of voluntary communication by supervisors and recognition of subordinates. Pelz (1951) from some of the same studies stresses the "power" of a supervisor, that it, how influential he may be with his own superiors. All of these have shown some relation to productivity. One later study from the Survey Research Center groups by Indik, et. al. (1961) has attempted to study some of these findings on the basis of four hypotheses. These concerned the enhancement of job performance by opinions of superior-subordinate communication, supportive behavior by superiors, mutual understanding among members and feelings of influence over local operations. With four criteria of recorded production, "station" production, and ratings of individual effectiveness and station effectiveness generally, positive associations were found in all tests for the organization as a whole and stations as such; however, analysis of individual stations gave widely varying results. Here again only a one-month time period was covered. A longer time might have given more opportunity for relationships to become pronounced.

The Southern California Organization Research Project published a series of studies which generally supported the findings

of the Survey Research Center. However, two of the studies dealing with skilled craftsmen at the San Diego Naval Air Station, (Wilson, et. al., 1953 and 1954), found supervisors of high and low producing groups similar in the first study, and with no differences in a second. These studies have also found the existence of curvilinear relationships, effective supervisors have more confidence in their subordinates both in personal and performance aspects and they have, in the more recent investigations, tended to become critical of psychologists' emphasis upon the interpersonal as contrasted with the technical aspects of supervision.

In summary, the investigations of the effects of supervision and/or organizational characteristics seem to indicate some rather modest effect. However, negative findings or specificity always create a nagging doubt--Is supervision a moderator?

The question of the influence of situational variables seems to indicate, from the presented material, that there is some modification of job performance by such variables. However, the general conception of studies with one independent and one dependent variable has led to the situation where modest relationships, contradictory results or no results at all have become commonplace. It would seem that studies such as those by Stodgill, et. al. (1955), and Wherry, et. al. (1961), both previously discussed, are the immediate need. Reported studies have indicated specifics to be looked for and evaluated for their effects on job performance, but experimental investi-

gations of entire organizations with a gradual working down toward sub-units and individuals must be conducted before the parameters of organizational effects can be established with any degree of confidence.

Conclusions and Hypotheses

From the foregoing review, it is apparent that job performance variance has been shown or presumed to be a result of a wide range of causal influences and its measurement is nebulous. In general, the authors submit, that if any significant progress is to be made toward solution of these problems, some basic research conceptions will need to be recast and broadened.

As a possible starting point, it is suggested that job performance will need to be viewed as both an independent and dependent variable having measurable outputs resulting from the interaction of job behaviors, situation characteristics, and personal characteristics. Broadly these outputs might be thought of as economic, adjustment, and personal. The first is measured by production indices, the second by reports such as absenteeism, and the last by survey techniques or, in conjunction with certain objective indices as grievances or disciplinary actions. If such indices can be established, it should be possible to design more complex studies encompassing organization or broad sub-units to determine interdependence, relative importance and, most importantly, causal bases of various dimensions of job performance. In addition, organizational indices do measure something "real" in contrast to global ratings that seem to bear little or no relationship to objectively measured performance. Such indices may not be the most desirable from the viewpoint of statistical evaluation, for example skewness, but from a strictly pragmatic stand, they

exist and possibly represent the only avenue for working toward a better understanding of job performance and criteria.

To define the starting point and guide succeeding investigations some seemingly fruitful hypotheses are suggested below.

Longitudinal studies (five or more years) will allow much better predictions of performance than shorter studies.

Performance indices of the organization or sub-units are required before complete assessment of individual performance can be accomplished. A sub-hypothesis would be: economic and "satisfaction producing" effects of job behavior are bi-polar.

Use of "organizational indices" as absenteeism, accidents, production, scrap, turnover, etc., as criteria will yield "purer" more predictable criteria of job performance, tending toward orthogonality. A sub-hypothesis would be: bi-polar interrelations will be found, in particular, at higher job levels.

Organizational indices will reveal common performance factors for functionally similar jobs with different patterns of success or failure for functionally different jobs.

Increasing required levels of performance for particular jobs will result in higher performance reliabilities and validities. A sub-hypothesis would be: specifically, more individual liberty to "do the job" will result in better job performance.

"Human evaluation" of performance will show low relationships with objective indices of performance. Some sub-hypotheses would be: some performances can only be evaluated by

human observation, i.e., Heron's (1952) op. cit., "source of concern to their supervisors."; performance evaluation ability is a predictable individual difference; reliability of judgmental criteria will vary inversely with proximity, in particular, peer evaluation will provide better criteria than supervisory.

Different predictors and criteria are more appropriate at different points in time, i.e., training vs. on-the-job, younger vs. older employees.

Predictor, individual, situational and organizational patterns of sub-groups and interrelations will be revealed by splitting criterion groups into halves, thirds, or some other sections. A sub-hypothesis would be: now unconceived hypotheses will be uncovered by the major hypothesis.

Job performance variability reliability is a predictable individual characteristic, e.g., classic reliability theory does not apply to individual performance. Some sub-hypotheses would be: unless performance reliability is held constant, group validities will remain low; performance complexity is inversely related to reliability; performance reliability is a probable job performance criterion; situational moderator variables may inflate or restrict reliabilities.

Certain performances, "creativity" for example, will show close to zero reliabilities.

Measured performance reliability will increase as a function of (A) time span for measuring increases and (B) purer criterion measures.

Measures of individual satisfaction as criteria will add another, separate, dimension to job performance. A sub-hypothesis would be: job satisfaction is an expression of a more deeply based general satisfaction.

Job performance variance, resulting from morale and attitudes, is comparatively small. Some sub-hypotheses would be: morale and attitudes function as moderator or mediating variables and do not directly affect performance; half the variance from the major hypothesis is specific; sources of this variance affect job performance differently at different levels and different jobs.

Moderator variables, most as yet untested, will have to be isolated to determine differential effects on predictor-performance relationships. Some sub-hypotheses would be: such effects will be substantial in the case of basic differences, as sex, for different jobs; functional job analysis will reveal conflicts in job composition, the conflict resulting from opposing requirements of person characteristics; functional situational analysis will reveal moderator variables heretofore defined as job performance variables of jobs in higher, the same, or lower levels in the organization hierarchy.

Classical statistical techniques will give way to some form of pattern analysis in analyzing and predicting job performance.

Whether or not the above hypotheses are adequate it is apparent that some drastic research approach is required if any progress is to be made in personnel research. Ghiselli's

review (1955) showed that little progress had been made after approximately 40 years of research effort. However, some recent studies have made promising beginnings in the direction indicated by the above hypotheses.

In measurement, Weitz (1961), op. cit., shows that conclusions in our experiments are dependent upon the criterion employed. Guion (1961), op. cit., and Dunnette (1963b), op. cit., have proposed modifications of the conventional approach to criterion utilization that have wide implications for criterion measurement. Fiske (1951) has discussed criteria and suggested defining and measuring job functions and using as criteria their contribution to the successful functioning of lower echelons in an organization. Stark (1959) has made much the same point, limited to executive success, in that executive jobs would be classified according to functions as supervising, planning, negotiating, investigating or some combination of these. On a more limited basis, Enell and Haas (1960) and Patton (1950) discuss evaluation of executive performance in terms of, in the former, comparing sub-unit targets implicit in a particular job. Both would then arrive at an evaluation of executive performance based upon comparing unit performance with goals set by either method. Lamouria and Harrell (1963), compensate for differences in the importance of company objectives in 19 different departments. Differences in functions were evaluated objectively and clinically and the resultant criterion scores for individuals (and departments) were judged to be less contaminated than are clinical ratings. In effect, all of these

studies, more or less explicitly, recognize that job performance has an outcome, the outcome can be evaluated in and of itself or against some standard and, implicitly, the need for broader, objective performance measures.

Toops (1959), op. cit., has called attention to many of the points made in this review, but the number of studies attempting to follow his suggestions has been limited.

Studies previously cited, McQuitty, et. al. (1954), Seashore, et. al. (1960). Wherry, et. al. (1961), and Stogdill, et. al. (1955), have studied job performance in the much broader context of organizational setting and such studies seem to be the desirable direction of personnel research in order to overcome the generally disappointing results obtained in more limited studies or, possibly, to make a new beginning. The designs used in these studies show the way toward models which might begin a more intensive investigation of the variables that do or might affect job performance. The complexities of such studies have been discussed by Dunnette (1963b), op. cit., using the Guetzkow and Forehand (1961) model and they are formidable but with modern computers the possibility of isolating job performance bases seems to be more promising. However, the question arises as to what to measure, how to measure and, perhaps most important of all, can reliable measures be made? That these questions are pertinent is indicated by the 1954 McQuitty study where, with a "descriptive inventory" of 264 items, 23 factors were extracted which accounted for only slightly over 50% of the variance. The generally negative results of the Seashore

study have already been commented upon. Both the Stogdill and Wherry studies were more encouraging, but it appears some basic hypotheses must be evaluated before much further progress is possible even with broadened research designs.

Probably the most important consideration is an abandonment of global criteria. As Dunnette (1963), op. cit. has pointed out, over-simplified studies consistently have ignored the many facets of job success and, in light of the studies discussed in the second section of this review, there can be no question of the multidimensional nature of even the simplest job.

However, even these studies do not seem to be broad enough in scope or time to solve the "criterion problem." It is suggested that future investigations must be conceived on a much broader scale in order to answer the questions posed in the separate sections of thie review.

References

Adjutant General's Office, Personnel Research in the Army, VI. The selection of truck drivers. Psychological Bulletin, 1943, 40, 499-508.

Adkins, Dorothy C. Construction and analysis of achievement tests. U. S. Government Printing Office, Washington, D. C., 1947.

Aiken, L. R. The grading behavior of a college faculty. Educational and Psychological Measurement, 1963, 23, 319-322.

Anastasi, Anne. Practice and variability. A study in psychological method. Psychological Monographs: General and Applied, 1934, 45, 5.

Anastasi, Anne. The concept of validity in the interpretation of test scores. Educational and Psychological Measurement, 1950, 10, 67-77.

Ayers, A. W. A comparison of certain visual factors with the efficiency of textile inspectors. Journal of Applied Psychology, 1942, 26, 812-827.

Baier, D. E., & Dugan, R. D. Factors in sales success. Journal of Applied Psychology, 1957, 41, 37-40.

Bair, J. T., Lockman, R. F., & Martoccia, C. T. Validity and factor analyses of naval air training, predictor and criterion measures. Journal of Applied Psychology, 1956, 40, 213-219.

Ball, R. S. The predictability of occupational level from intelligence. Journal of Consulting Psychology, 1938, 2, 184-186.

Balma, M. J., Ghiselli, E. E., McCormick, E. J., Primoff, E. S., & Griffin, C. H. The development of processes for indirect or synthetic validity, (a symposium). Personnel Psychology, 1959, 12, 395-420.

Bartelme, Phyllis F., Fletcher, E. D., Brown, C. W., & Ghiselli, E. E. The prediction of driving skill. Journal of Applied Psychology, 1951, 35, 98-100.

Bass, B. M. Further evidence on the dynamic character of criteria. Personnel Psychology, 1962, 15, 93-97.

Bayroff, A. G., Haggerty, Helen R., & Rundquist, E. A. Validity of ratings as related to rating techniques and conditions, Personnel Psychology, 1954, 7, 93-113.

Bechtoldt, H. P. Selection. In S. S. Stevens (Ed.), Handbook of experimental psychology. New York: John Wiley & Sons, 1951.

Bellows, R. M. Studies of clerical workers, Chapter VIII in Occupational Counseling Techniques. Stead, C. L., Shartle, C. L., et. al. (Eds.) New York: American Book Company, 1940.

Bellows, R. M. Procedures for evaluating vocational criteria. Journal of Applied Psychology, 1941, 25, 499-513.

Bird, N. Relationships between experience factors, test scores and efficiency. Archives of Psychology, New York, No. 126, 1931.

Bockner, D. N. The prediction of ratings as a function of interrater agreement. Journal of Applied Psychology, 1946, 30, 23-31.

Bolanovich, D. J. Statistical analysis of an industrial rating chart. Journal of Applied Psychology, 1946, 30, 23031.

Braunhaussen, N. Selection des employees de bureau. Revue de la science du travail, 1929, 1, 499-512.

Brayfield, A. H., & Crockett, W. H. Employee attitudes and performance. Psychological Bulletin, 1955, 52, 296-424.

Brogden, H. E., & Taylor, E. K. The theory and classification of criteria bias. Educational and Psychological Measurement, 1950, 10, 159-186. (b)

Campbell, D. Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 1957, 54, 4, 297-312.

Carter, L. F., & Dudek, F. J. The use of psychological techniques on measuring and critically analyzing navigators' flight performance. Psychometrika, 1947, 12, 31-42.

Clark, K. E. The vocational interests of nonprofessional men. Minneapolis: University of Minnesota Press, 1961.

Clarke, F. R. Labor turnover studies. Personnel Journal, 1946, 25, 55-58.

Cleven, W. A., & Fieldler, F. E. Interpersonal perceptions of open hearth foremen and steel production. Journal of Applied psychology, 1956, 40, 312-314.

Cohen, L., & Strauss, L. Time study and the fundamental nature of manual skill. Journal of Consulting Psychology, 1946, 10, 146-153.

Comrey, A. L. A factorial study of achievement in West Point courses. Educational and Psychological Measurement, 1949, 9, 193-209.

Coombs, C. H. Some hypotheses for the analysis of qualitative variables. Psychological Review, 1948, 55, 167-174.

Craig, D. R. The preference--interest questionnaire in selecting retail saleswomen. Journal of Personnel Research. 1924-1925, 3, 366-374.

Cureton, E. E. Note on the validity of the American Council on Education Psychological Examination. Journal of Applied Psychology, 1939, 23, 306-307.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Cureton, E. E., & Katzell, R. A. A further analysis of the relations among job performance and situational variables. Journal of Applied Psychology, 1962, 46, 230.

Dalton, M. Informal factors in career achievement. American Journal of Sociology, 1951, 56, 407-415.

Davis, J. A. Faculty definition of desirable student traits. Faculty Perception of Students, Research Bulletin RB-64-11, Educational Testing Service, Princeton, N. J., 1964. (a)

Davis, J. A. Structure of faculty characterizations. Faculty Perception of Students, Research Bulletin RB-64-12, Educational Testing Service, Princeton, N. J., 1964. (b)

Davis, J. A. Desirability and perception of academic performance. Faculty Perception of Students, Research Bulletin RB-64-13, Educational Testing Service, Princeton, N. J., 1964. (c)

Dorcus, R. M. Methods of evaluating the efficiency of door-to-door salesmen of baking products. Journal of Applied Psychology, 1940, 24, 587-594.

Dudek, E. E. Personnel selection. In Paul Farnsworth (Ed.), Annual review of psychology, 14, 1963.

Dudek, F. J. The dependence of factorial composition of aptitude tests upon the population differences among pilot trainees II. The factorial composition of test and criterion variables. Educational and Psychological Measurement, 1949, 9, 95-104.

Dunnette, M. D. A note on the criterion. Journal of Applied Psychology, 1963, 47, 251-253. (a)

Dunnette, M. D. A modified model for test validation and selection research. Journal of Applied Psychology, 1963, 47, 317-332. (b)

Eckerman, A. C. An analysis of grievances and aggrieved employees in a machine shop and foundry. Journal of Applied Psychology, 1948, 32, 255-269.

Edgerton, H. A., & Kolbe, L. E. The method of minimum variation for the combination of criteria. Psychometrika, 1936, 1, 183-187.

Enell, J. W., & Haas, G. H. Setting standards for executives performance. American Management Association, New York, 1960, Research Study 42. Standard Oil (Ohio). Standards of performance.

Ewart, E., Seashore, S. E., & Tiffin, J. A factor analysis of an industrial merit rating scale. Journal of Applied Psychology, 1941, 25, 481-486.

Eysenck, H. J. The measurement of motivation. Scientific American, 1963, 208, 130-140.

Fay, P. J., & Middleton, W. C. Relationship between sales ability and ratings of the transcribed voices of salesmen. Journal of Applied Psychology, 1942, 26, 499-510.

Ferguson, L. W. Management quality and its effect on selection test validity. Personnel Psychology, 1951, 4, 141-150.

Ferguson, L. W. Ability, interest and aptitude. Journal of Applied Psychology, 1960, 44, 126-131.

Fiske, D. W. Values, theory and the criterion problem. Personnel Psychology, 1951, 4, 93-98.

Fiske, D. W. The constants of intra-individual variability in test response. Educational and Psychological Measurement, 1957, 17, 317-337. (a)

Fiske, D. W. An intensive study of variability scores. Educational and Psychological Measurement, 1957, 17, 453-465. (b)

Fiske, D. W., & Rice, Laura. Intra-individual response variability. Psychological Bulletin, 1955, 52, 217-250.

Flanagan, J. C. The experimental evaluation of a selection procedure. Educational and Psychological Measurement, 1946, 6, 445-446.

Flanagan, J. C. (Ed.) The aviation psychology program in the Army Air Force. Army Air Forces Aviation Psychology Program Research Report, 1. Washington: U. S. Government Printing Office, 1948.

Flanagan, J. C. Job requirements. In W. Dennis (Ed.), Current trends in industrial psychology. University of Pittsburgh, 1949, 32-54.

Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358. (a)

Flanagan, J. C. Job element aptitude classification tests, Personnel Psychology, 1954, 7, 1-14. (b)

Flanagan, J. C. The evaluation of methods in applied psychology and the problem of criteria. Occupational Psychology, 1956, 30, 1-9.

Fleishman, E. A., & Fruchter, B. Factor structure and predictability of successive stages of learning Morse Code. Journal of Applied Psychology, 1960, 44, 97-101.

Fleishman, E. A., Harris, E. F., & Burtt, H. E. Leadership and supervision in industry. Bureau of Educational Research Monograph No. 33. Columbus, Ohio: The Ohio State University, 1955.

Fleishman, E. A., & Ornstein, G. N. An analysis of pilot flying performance in terms of component abilities. Journal of Applied Psychology, 1960, 44, 146-155.

Frey, D. M. Selection of promotion salesmen. Journal of Personnel Research. 1925-26, 5, 142-156.

Freyd, M. Measurement in vocational selection. Journal of Personnel Research, 1923-24, 2, 215-249.

Gadel, Marguerite S., & Kreidt, P. H. Relationships of aptitude, interest, performance and job satisfaction of IBM operators. Personnel Psychology, 1952, 5, 207-212.

Gaier, E. L. The criterion problem in the prediction of medical school success. Journal of Applied Psychology, 1952, 36, 316-322.

Gaylord, R., Russell, Eva, Johnson, C., & Severen, D. The relation of ratings to production records: An empirical study. Personnel Psychology, 1951, 4, 363-371.

Ghiselli, E. E. The measurement of occupational aptitude. Berkeley: University of California Press, 1955.

Ghiselli, E. E. The placement of workers: concepts and problems. Personnel Psychology, 1956, 9, 1-16.

Ghiselli, E. E. The prediction of predictability. Educational and Psychological Measurement, 1960, 20, 3-8. (a)

Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. Educational and Psychological Measurement, 1960, 20, 675-684. (b)

Ghiselli, E. E. Moderating effects and differential reliability and validity. Journal of Applied Psychology, 1963, 47, 81-86.

Ghiselli, E. E. & Barthol, R. P. The validity of personality inventories in the selection of employees. Journal of Applied Psychology, 1953, 37, 18-20.

Ghiselli, E. E., & Barthol, R. P. Role perceptions of successful and unsuccessful supervisors. Journal of Applied Psychology, 1956, 40, 241-244.

Ghiselli, E. E., & Brown C. W. Validity of aptitude tests for predicting trainability of workers. Personnel Psychology, 1951, 4, 243-260.

Ghiselli, E. E., & Haire, M. The validation of selection tests in the light of the dynamic nature of criteria. Personnel Psychology, 1960, 13, 225-231.

Gilmer, V. H. Industrial psychology. New York: McGraw-Hill, 1961.

Gottsdanker, R. M. Measures of potentiality for machine calculation. Journal of Applied Psychology, 1943, 27, 233-248.

Graham, W. R. Identification and prediction of two training criterion factors. Journal of Applied Psychology, 1954, 38, 96-99.

Grant, D. L. A factor analysis of managers' ratings. Journal of Applied Psychology, 1955, 39, 283-286.

Greystoke, J. R., Thomason, C. F., & Murphy, J. J. Labor turnover surveys. Journal of the Institute of Personnel Management, 1952, 34, 158-165.

Guetzkow, H., & Forehand, G. A. A research strategy for partial knowledge useful in the selection of executives. In R. Taguiri (Ed.), Research needs in executive selection. Boston: Harvard Graduate School of Business Administration, 1961.

Guilford, J. P. (Ed.) Printed classification tests. USAF Report No. 5, Army Air Force Aviation Psychology Research Reports. Washington: U. S. Government Printing Office, 1947.

Guilford, J. P. Three faces of intellect. The Walter V. Bingham Memorial Lecture, Stanford University, April 13, 1959.

Guion, R. M. Criterion measurement and personnel judgments. Personnel Psychology, 1961, 14, 141-149.

Haggerty, Helen R., Johnson, C. D., & King, S. H. Evaluation of mail-order ratings on combat performance of officers. Personnel Psychology, 1959, 12, 597-205.

Haire, M. Psychological problems relevant to business and industry. Psychological Bulletin, 1951, 56, 3, 169-194.

Hay, E. N. Predicting success in machine bookkeeping. Journal of Applied Psychology, 1943, 27, 483-493.

Hayes, E. G. Selecting women for shop work. Personnel Journal, 1932-33, 11, 69-85.

Hemphill, K. J. Dimensions of executive positions. Research Monograph, No. 98, Bureau of Business Research, Ohio State University, 1960.

Henry, W. E. The business executive: The psychodynamics of a social role. American Journal of Sociology, 1949, 54, 286-291.

Heron, A. A psychological study of occupational adjustment. Journal of Applied Psychology, 1952, 36, 385-387.

Heron, A. Satisfaction and satisfactoriness. Complementary aspects of occupational adjustment. Occupational Psychology, 1954, 28, 140-153.

Hertzman, M. The influence of the individual's variability on the organization of performance. Journal of General Psychology, 1939, 20, 3-24.

Herzberg, F., Mausner, B., Peterson, R. O., & Capwell, Dora F. Job attitudes - review of opinion and research. Psychological Service of Pittsburgh, 1957.

Herzberg, F., Mausner, B., & Snyderman, Barbara B. The motivation to work. New York: John Wiley & Sons, 1959.

Hicks, J. A., & Stone, J. B. The identification of traits related to managerial success. Journal of Applied Psychology, 1962, 46, 428-432.

Hilton, T. L., & Dill, W. R. Salary growth as a criterion of career progress. Journal of Applied Psychology, 1962, 46, 153-158.

Hollander, E. P. Buddy ratings: Military research and industrial implications. Personnel Psychology, 1954, 7, 385-398.

Hollander, E. P. Interpersonal exposure time as a determinant of the predictive utility of peer ratings. Psychological Reports, 1956, 2, 445-448.

Hollander, E. P., & Webb, W. B. Leadership, followership and friendship; an analysis of peer nominations. Journal of Abnormal and Social Psychology, 1955, 50, 163-167.

Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. Psychometrika, 1936, 1, 53-60.

Horst, P. (Ed.) The prediction of personal adjustment. New York: Social Service Research Council, 1941.

Hotelling, H. The most predictable criterion. Journal of Educational Psychology, 1935, 26, 139-142.

Hotelling, H. Relations between two sets of variates. Biometrika, 1936, 28, 321-377.

Hull, C. L. Aptitude testing. New York: World Book Company, 1928.

Huse, E., & Taylor, E. Reliability of absence measures. Journal of Applied Psychology, 1962, 46, 159-160.

Indik, B. P., Georgopoulos, B. S., & Seashore, S. E. Superior-subordinate relationships and performance. Personnel Psychology, 1961, 14, 357-374.

Jaspen, N. A factor study of worker characteristics. Journal of Applied Psychology, 1949, 33, 449-459.

Jenkins, J. G. Validity for what? Journal of Consulting Psychology, 1946, 10, 93-98.

Kahn, R. L., & Morse, Nancy C. The relation of productivity to morale. Industrial Relations Research Association, Proceedings of Fourth Annual Meeting, Boston, Mass., December 28-29, 1951, 69-85.

Katzell, R. A. Contrasting systems of work organization. American Psychologist, 1962, 17, 102-108.

Keenan, V., Kerr, W. A., & Sherman, W. Psychological climate and accidents in an automotive plant. Journal of Applied Psychology, 1951, 35, 108-111.

Kellner, A. D. The use of interim measures of performance and suppressor variables in appraising employee potential. Journal of General Psychology, 1960, 62, 19-23.

Kerber, H. E. A study of individual differences in on-the-job behavior and inquiry. Unpublished doctoral dissertation, Western Reserve University, Cleveland, 1958.

Kerr, W. A. Accident proneness of factory departments. Journal of Applied Psychology, 1950, 34, 167-170.

Kettner, N. W., Guilford, J. P., & Christensen, P. R. The relation of certain thinking factors to training criteria in the U. S. Coast Guard Academy, Educational and Psychological Measurement, 1959, 19, 381-394.

King, D. C. A multiplant factor analysis of employees' attitudes toward their company. Journal of Applied Psychology, 1960, 44, 241-243.

Kingsbury, F. A. Making rating scales work. Journal of Personnel Research, 1925-26, 4, 1-6.

Kingsbury, F. A. Psychological tests for executives. Personnel, 1933, 9, 121-133.

Kipnis, D. Some determinants of supervisory esteem. Personnel Psychology, 1960, 13, 377-391.

Kipnis, D. A non-cognitive correlate of performance among lower aptitude men. Journal of Applied Psychology, 1962, 46, 76-80.

Kirchner, W. K. Predicting ratings of sales success with objective performance information. Journal of Applied Psychology, 1960, 44, 398-403.

Klemmer, E. T., & Lockhead, G. R. Productivity and errors in two keying tasks: A field study. Journal of Applied Psychology, 1962, 46, 401-408.

Knauft, E. B. Classification and evaluation of personnel rating methods. Journal of Applied Psychology, 1947, 31, 617-625.

Kornhauser, A. W. A statistical study of a group of specialized office workers. Journal of Personnel Research, 1923-24, 2, 103-123.

Kossoris, M. D. Relation of age to industrial injuries. Monthly Labor Review, 1940, 51, 789-804.

Lamouria, L. H., & Harrell, T. W. An approach to an objective criterion for research managers. Journal of Applied Psychology, 1963, 47, 353-357.

Langdon, J. N. An experimental study of certain forms of manual dexterity. Great Britain Medical Research Council, Industrial Health Research Board, Report No. 66, 1932, 56.

Lawshe, C. H., & McGinley, A. D. Job performance criteria studies: The job performance of proof readers. Journal of Applied Psychology, 1951, 35, 316-320.

Lawshe, C. H., & Steinberg, M. D. Studies in synthetic validity: An exploratory investigation of clerical jobs. Personnel Psychology, 1955, 8, 291-301.

Lifson, K. A. Errors in time-study judgments of industrial work pace. Psychological Monographs: General and Applied, 1953, 67, No. 355.

Likert, R. Findings of research on management and leadership. Proceedings, Pacific Gas Association, 1951, 43.

Link, H. C. Employment Psychology. New York: The MacMillan Company, 1919.

Liske, R. E., Ort, R. S., & Ford, Amasa B. Clinical Performance ratings of medical students and faculty physicians. Report Under Research Grant No. 607 and 1060-P Health, Education, and Welfare, Washington, D. C.

Locke, E. A. The development of criteria of student achievement. Educational and Psychological Measurement, 1963, 23, 299-308.

Lovett, R. F. A study of the application blanks, service records and production records of 1129 salesmen of the Proctor & Gamble Company, M. A. Thesis, Carnegie Institute of Technology, 1923.

Lurie, W. A. The concept of occupational adjustment. Educational and Psychological Measurement, 1942, 2, 3-14.

Maccoby, N. Research findings on productivity supervision, and morale. Research on Human Relations in Administration, Unpublished report, University of Michigan, Institute for Social Research, 1949.

Mackie, R. R., & High, W. S. Supervisory ratings and practical performance tests as complementary criteria of shipboard performance. Los Angeles, California: Human Factors Research, 1959.

MacKinney, A. C., & Wolins, L. Validity information exchange. Personnel Psychology, 1960, 13, 443-447.

Mann, F. C., & Dent, J. Appraisals of supervisors and attitudes of their employees in an electric power plant. Ann Arbor, Michigan: University of Michigan, Survey Research Center, 1954.

Manning, W. H., & DuBois, P. H. Gain in proficiency as a criterion in test validation. Journal of Applied Psychology, 1958, 42, 191-194.

Marriott, R. Size of working group and output. Occupational Psychology, 1949, 23, 47-57.

Matthews, J. Research on development of valid situational tests of leadership. I. Survey of the literature. Pittsburgh, Pennsylvania: American Institute for Research, 1951.

McQuitty, L. L., Wrigley, C., & Gaier, E. K. An approach to isolating dimensions of job success. Journal of Applied Psychology, 1954, 38, 227-232.

Merrihue, M. & Katzell, R. A. ERI-Yardstick of employee relations. Harvard Business Review, 1955, 33, (6), 91-99.

Michael, W. B. Factor analyses of tests and criteria. Psychological Monographs: General and Applied, 1949, 63, 298.

Miner, J. B. Personality and ability factors in sales performance. Journal of Applied Psychology, 1962, 46, 6-13.

Mintz, A. Time intervals between accidents. Journal of Applied Psychology, 1954, 38, 6.

Nagle, B. F. Criterion development. Personnel Psychology, 1953, 6, 271-288.

Newman, S. H., French, J. W., & Bobbitt, J. M. Analysis of criteria for the validation of selection measures at the United States Coast Guard Academy. Educational and Psychological Measurement, 1952, 12, 394-407.

Otis, J. L. Whose criterion? Presidential Address to Division 14, American Psychological Association Convention, 1953.

Otis, J. L., Endler, O. L., & Kolbe, L. E. Data-analysis methods. In W. H. Stead, et. al. (Eds.), Occupational counseling techniques. New York: American Book Company, 1940.

Owens, W. A. Intra-individual differences versus inter-individual differences in motor skills. Educational and Psychological Measurement, 1942, 2, 301-314.

Palmer, G. J., & McCormick, E. J. A factor analysis of job activities. Journal of Applied Psychology, 1961, 45.

Palmer, G. J. & Schroeder, R. H. Incentive conditions and behavior in 188 Industrial Manufacturing Organizations. Technical Report No. 3, Office of Naval Research, 1961.

Parkinson, R. Turnover and length of service of salesmen. Personnel, 1928, 1, 155-161.

Patton, A. How to appraise executive performance: Planned performance. Harvard Business Review, 1960, 38, 63-70.

Peck, R. F., & Parsons, J. W. Personality factors in work output; Four studies of factory workers. Personnel Psychology, 1956, 9, 49-79.

Pelz, D. Leadership within a hierarchial organization. Journal of Social Issues, 1951, 7, 49-55.

Penn, R. An investigation of some of the methodological problems concerned with accident proneness research. Unpublished doctoral dissertation. Carnegie Institute of Technology, 1955.

Peres, S. Performance dimensions of supervisory positions. Personnel Psychology, 1962, 15, 405-410.

Peters, R., & Campbell, J. T. Diagnosis of training needs of B-29 mechanics from supervisory ratings and self-ratings. Technical memorandum, 1955, Personnel Research Laboratory-Technical Manual-55-12.

Pond, Millicent. Selective placement of workers. Journal of Personnel Research, 1925-26, 5, 345-368.

Prien, E. P. Development of a supervisor position description questionnaire, Journal of Applied Psychology, 1963, 47, 10-14.

Prien, E. P., & Lee R. Analysis of ten criteria of student performance. Psychological Reports, 1965, 17, 273-274.

Prien, E. P., & Powell, D. R. A study of the director's functions. Journal of the American Society of Training Directors, 1961, 15, 12-17.

Primoff, E. W. The coefficient approach to jobs and tests. Personnel Administration, 1957, 20, No. 3.

Richards, J. M., Taylor, C. W., Price, D. D., & Jacobsen, T. L. Investigation of the criterion problem for one group of medical specialists. Journal of Applied Psychology, 1965, 49, 79-90.

Roach, D. E. Factor analysis of rated supervisory behavior. Personnel Psychology, 1956, 9, 487-498.

Ronan, W. W. A factor analysis of eleven job performance measures. Personnel Psychology, 1963, 16, 255-267.

Rothe, H. F. Output rates among butter wrappers: I. Work curves and their stability. Journal of Applied Psychology, 1946, 30, 199-211. (a)

Rothe, H. F. Output rates among butter wrappers: II. Frequency distributions and an hypothesis regarding the "Restriction of output." Journal of Applied Psychology, 1946, 30, 320-328. (b)

Rothe, H. F. Output rates among machine operators: I. Distributions and their reliability. Journal of Applied Psychology, 1947, 31, 484-489.

Rothe, H. F. Output rates among chocolate dippers. Journal of Applied Psychology, 1951, 35, 94-97.

Rothe, H. F., & Nye, C. T. Output rates among coil winders. Journal of Applied Psychology, 1958, 42, 182-186.

Rothe, H. F., & Nye, C. T. Output rates among machine operators: II. Consistency related to methods of pay. Journal of Applied Psychology, 1959, 43, 417-420.

Rush, C. H., Jr. A factorial study of sales criteria. Personnel Psychology, 1953, 6, 9-24.

Ryans, D. G., & Frederiksen, N. Performance tests of educational achievement. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Sawatsky, J. C. Psychological factors in industrial organizations affecting employee stability. Canadian Journal of Psychology, 1951, 5, 29-38.

Seashore, S. E. The aptitude hypotheses in motor skills. Journal of Experimental Psychology, 1931, 14, 555-561.

Seashore, S. H., Indik, B. P., & Georgopoulos, B. S. Relationships among criteria of job performance. Journal of Applied Psychology, 1960, 44, 195-202.

Shellow, Sadie M. Selection of motormen: Further data on value of tests in Milwaukee. Journal of Personnel Research, 1925-26, 5, 183-188.

Siegel, A. I. The check list as a criterion of proficiency. Journal of Applied Psychology, 1954, 38, 93-95.

Siegel, A. I., Schultz, D. G., & Benson, S. Post-training performance criterion development and application. Wayne, Pa.: Applied Psychological Services, March, 1960.

Smith, Patricia C., & Gold, R. A. Prediction of success from examination of performance during the testing period. Journal of Applied Psychology, 1953, 37, 69-74.

Sprecher, T. B. A study of engineers' criteria for creativity. Journal of Applied Psychology, 1959, 43, 141-148.

Springer, Doris. Rating of candidates for promotion by co-workers and supervisors. Journal of Applied Psychology, 1953, 37, 347-351.

Stark, S. Research criteria of executive success. Journal of Business, 1959, 32, 1-14.

Stead, W. H., Shartle, C. L., et. al. (Eds.). Occupational counseling techniques. New York: American Book Company 1940.

Stogdill, R. M., Shartle, C. L., Wherry, R. J., & Jaynes, W. E. A factorial analytic study of administrative behavior. Personnel Psychology, 1955, 8, 165-180.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, Shirley, & Clausen, J. C. Measurement and prediction. Studies in Social Psychology in World War II, Vol. 4. Princeton, N. J.: Princeton University Press, 1950.

Strong, E. K., Jr. Interest and sales ability. Personnel Journal, 1934-35, 13, 204-216.

Strong, E. K., Jr. Vocational interests of men and women. California: Stanford University Press, 1943.

Stuit, D. B., & Wilson, J. T. The effect of an increasingly well defined criterion on the prediction of success at naval training school (tactical radar). Journal of Applied Psychology, 1946, 30, 614-623.

Stuit, D. B. (Ed.) Personnel research and test development in the Bureau of Naval Personnel. Princeton, N. J.: Princeton University Press, 1947.

Svetlik, B. A study of the relationship of expressed attitudes about one's job and job conditions to personality variables as measured by various personality tests. Unpublished M. A. Thesis, Western Reserve University, Cleveland, Ohio, 1961.

Taylor, C. W. The Third (1959) University of Utah Research Conference on the Identification of Creative Talent, 1959.

Taylor, E. K., & Munson, Grace E. Supervised ratings--making graphic scales work. Personnel, 1951, 27, 504-514.

Taylor, E. K., Munson, Grace E., & Stone, P. M. Validation of tests for routine I.B.M. jobs at Office of Dependency Benefits, Newark, New Jersey, PRS Report No. 698. Personnel Research Section, the Adjutant General's Office, War Department, 1945.

Thorndike, E. L. A constant error in psychological ratings. Journal of Applied Psychology, 1920, 4, 25-29.

Thorndike, R. L. Personnel selection. New York: Wiley, 1949.

Toops, H. A. The criterion. Educational and Psychological Measurement, 1944, 4, 271-297.

Toops, H. A. A research utopia in industrial psychology. Personnel Psychology, 1959, 12, 189-225.

Travers, R. M. N. The use of a discriminant function in the treatment of psychological group differences. Psychometrika, 1939, 4, 25-32.

Turner, W. W. Dimensions of foreman performance: A factor analysis of criterion measures. Journal of Applied Psychology, 1960, 44, 216-223.

Van Dusen, A. C. Importance of criteria in selection and training. Educational and Psychological Measurement, 1947, 7, 498-504.

Viteles, M. S. Standards of Accomplishment: Criteria of vocational selection. Journal of Personnel Research, 1926, 4, 483-486.

Viteles, M. S. A dynamic criterion. Occupations, 1936, 14, 963-967.

Walker, R. Y., Bennett, S. V., & Ewart, E. S. A study of individual differences among flight instructors in making spot landings. Washington, D. C.: CAA Division of Research Report #56, February 1946.

Wallace, S. R., Jr., & Weitz, J. Industrial psychology. Annual Review of Psychology, 1955, 6, 217-50.

Weitz, J. Criteria for criteria. American Psychologist, 1961, 16, 228-231.

Wells, F. L. A statistical study of literary merit. Archives of Psychology, New York, 1907, 1, No. 7.

Wherry, R. J. An approximation method for obtaining a maximized multiple criterion. Psychometrika, 1940, 5, 109-115.

Wherry, R. J. The past and future of criterion evaluation. Personnel Psychology, 1957, 10, 1-5.

Wherry, R. J., & Fryer, D. F. Buddy ratings: popularity contest or leadership criteria? Personnel Psychology, 1949, 2, 147-159.

Wherry, R. J., Stander, J., Leight, J., & Lecznar, W. B. General on-the-job criteria of airmen effectiveness applied to three career fields. Tech. Rep. ASD-TR, 61-98. Personnel Laboratory, U. S. Air Force, Lackland Air Force Base, Texas, 1961.

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

Whitlock, G. H. Application of the psychophysical law to performance evaluation. Journal of Applied Psychology, 1963, 47, 15-23.

Whitlock, G. H., Clouse, R. J., & Spencer, W. F. Predicting accident proneness. Personnel Psychology, 1963, 16, 35-44.

Wilson, R. C., Beem, Helen P., & Comrey, A. L. Factors influencing organizational effectiveness, III. A survey of skilled tradesmen. Personnel Psychology, 1953, 6, 313-325.

Wilson, R. C., High, W. S., Beem, Helen P., & Comrey, A. L. Factors influencing organizational effectiveness. IV. A survey of supervisors and workers. Personnel Psychology, 1954, 7, 525-531.

Yerkes, R. M. (Ed.) Psychological examining in the United States Army. Memoirs of the National Academy of Science. 1921, 15, 837.