in language translation: Technical English-to-Vietnamese. Research Paper P-634, Institute for Defense Analyses, Science and Technology Division, Arlington, Va., July, 1970.

Sinaiko, H. W., Guthrie, G. M., & Abbott, P. S. Operating and maintaining complex military equipment: A study of training problems in the Republic of Vietnam. Research Paper P-501, Institute for Defense Analyses, Science and Technology Division, Arlington, Va., July, 1969.

Sticht, T. G. Some relationships of mental aptitude, reading ability, and listening ability using normal and time-compressed speech. *The Journal of Communication*, 1968, 18, 243–258.

Sticht, T. G. Literacy demands of publications in selected military occupational specialities. Draft Professional Paper, Human Resources Research Office, Division Number 3 (Recruit Training), George Washington University, Washington, D.C., July, 1969.

Tannenbaum, A. S. *Social psychology of the work organization*. Belmont, Calif.: Wadsworth Publishing, 1966.

Teel, K. S. Is human factors engineering worth the investment? *Human Factors*, 1971, 13, 17–21.

Tiffin, J., & McCormick, E. J. *Industrial psychology*. (5th ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Weldon, R. J., & Peterson, G. M. Effect of design on accuracy and speed of operating dials. *Journal of Applied Psychology*, 1957, 41, 153–157.

Wilson, W. E. *Concepts of engineering system design*. New York: McGraw-Hill, 1965.

Woodrow, H. Time perception. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951, 1224–1236.

Woodson, W. E., & Conover, D. W. *Human engineering guide for equipment designers*. (2nd ed.) Berkeley: University of California Press, 1966.

CHAPTER 17

# Behaviors, Results, and Organizational Effectiveness: The Problem of Criteria[1]

PATRICIA C. SMITH
Bowling Green State University

In M.D. Dunnette (Ed.), Handbook of industrial and organizational Psychology. Chicago: Rand McNally, 1976.



DECIDING WHAT CRITERIA one wants to predict, manipulate, or conceptualize is the first problem of the scientist. Three dimensions seem to cover most criteria: the time-span to be covered, the specificity desired, and the closeness to organizational goals to be approached. These determine the logic of the criteria chosen, their relevance, reliability, and salability. Suggestions are given for improving criteria.

"The criterion" is certainly one of the key problems in industrial and organizational psychology, as evidenced by the massive efforts designed to clarify its theory and to improve its measurements. (For example, see such reviews as those of Kendall, 1956; Wallace & Weitz, 1955; Katzell, 1957; Taylor & Nevis, 1961; Weitz, 1961; Biesheuvel, 1965; Wallace, 1965; Guion, 1967; Owens & Jewell, 1969; Bray & Moses, 1972.) Determination of policies concerning organizational structure, managerial and employee development, conditions of work, design of jobs, incentive plans, leadership styles, selection, placement, transfer and promotion, and organizational objectives requires comparisons of criterion measures.

The dictionary (Funk & Wagnalls, 1963) defines a criterion as a "standard or rule by which a judgment can be made." In psychology it has come to mean a dependent or predicted measure for judging the effectiveness of persons, organizations, treatments, or predictors of behavior, results, and organizational effectiveness.

[1] I thank R. A. H. Goodfellow, O. W. Smith, J. Scheffers, and J. P. Wijting for their critical readings of the manuscript.

## REQUIREMENTS OF A CRITERION

The first requirement of a criterion is that it be relevant to some important goal of the individual, the organization, or society. Determination of relevance is, however, a matter of judgment. Some group or person must decide which activities are most relevant to success. Once these activities have been identified, efforts must then be directed toward developing psychometrically sound measures of these activities. The measure of a criterion should be neither contaminated with irrelevant variance nor deficient in terms of measuring the important objectives of the organization and of the people in it.

Neither the criterion nor the measure of it should be biased or trivial (Ghiselli & Brown, 1955, Chap. 4). In order to insure the importance of a criterion, careful analysis is required for the understanding of the goal or goals of the individual, organization, or society. Some goals are obvious. Some may be equally important, yet obscure and inadequately formulated. For the latter, criteria must be developed. These criteria are achieved only via analysis and understanding so that valid measures can be applied or devised. Consequently, relevancy, the first requirement for a criterion, consists of two parts. One is the validity of the goal which is judged to be important. The second is the validity of the measure(s) of the goal achievement. This requirement is parallel to the requirement that a test be valid.

Reliability is the second requirement of a criterion. It involves agreement between different evaluations, at different periods of time and with different although apparently similar measures. Reliability is usually said to set the upper limit of validity (Ryan & Smith, 1954, pp. 46-56; Wallace & Weitz, 1955; Blum & Naylor, 1968, p. 182) and of the predictability of effects of treatments. Reliability may, of course, be estimated or operationalized in several different ways, and these estimates may occasionally yield quite different values for the same criterion measure. For example, test-retest stability could be low even though internal consistency estimates derived at particular points in time may be high. Occasionally, therefore, and depending upon the type of "reliability" that is computed, the upper limit of validity may not be rigidly set by the value found for reliability. Dunnette (1966, pp. 35–36) discusses some anomalies wherein validity may not necessarily be limited by the size of reliability.

A criterion measure must in addition be practical—available, plausible, and acceptable to those who will want to use it for decisions. Psychologists are seldom the decision makers, and they must consider the market for their ideas. It is a temptation to predict that which is predictable (e.g., Kurtz, 1937), rather than that which ought to be predicted as a basis for decisions.

## MULTIDIMENSIONALITY OF CRITERIA

### The Search for the Criterion

The use of the single noun "criterion" has been misleading. It has stemmed from the false search for measures which would be related to what Thorndike (1949) called the "ultimate criterion."

> The ultimate criterion is the complete final goal of a particular type of selection or training. For example, it might have been agreed that the final goal in the selection and training of Air Force bombardiers was that they should under conditions of combat flying drop their bombs in every case with maximum precision upon the designated target. The ultimate goal in the selection and training of insurance salesmen might be that each man sell the maximum amount of insurance which would not be allowed to lapse and that he continue actively as an insurance salesman for an extended period of years. The ultimate criterion for a production line worker might be that he perform his task, maintaining the tempo of the line, with the minimum of defective products requiring rejection upon inspection, that he be personally satisfied with the task to such an extent that he is not a source of unrest and conflict with other workers, and that he continue in the job for an extended period of time. It can be seen that the ultimate goal is stated in very broad terms and in terms that are often not susceptible to practical quantitative evaluation. Furthermore, it is usually not entirely accurate to specify a single and unified ultimate goal. The bombardier had to fire a gun as well as drop bombs. The life insurance salesman must keep records and in many instances manage an office, as well as sell to customers. Even with the production line worker we have indicated considerations of contentment and permanence on the job as well as simple performance. A really complete ultimate criterion is multiple and complex in almost every case. Such a criterion is ultimate in the sense that we cannot look beyond it for any higher or further standard in terms of which to judge the outcomes of a particular personnel program.
>
> In practice, the complete ultimate criterion is rarely, if ever, available for use in psychological research. (Thorndike, 1949, p. 121)

As the quotation indicates, no single measure can fully express success or failure. Yet a single decision must somehow be made about each individual, organization, or treatment—hence the search for composites which properly represent the full complexity of the underlying measures.

There are several bases for combining measures into a single summary criterion measure. The first is statistical, based on intercorrelations of measures. If different measures are not highly correlated, then a composite is illogical, cancelling out important bits of information, pro and con. But if several measures can be shown to be highly intercorrelated or to represent a single factor, then construction of a combined measure is reasonable. An example of a statistical composite is given by Edgerton and Kolbe (1936), who obtain a composite by *maximizing* differences between individuals in terms of composite criterion scores and minimizing differences in scores on different criterion variables within the individual. Convergence of several criteria is illustrated by French (1954), who combined ratings from classmates, upperclassmen, and officers to obtain a single, reliable criterion of leadership. There are many other examples of successful combinations of criteria, but always to a rather limited goal.

The second basis is economic. If organizational goals can be reduced to common measurement scales such as dollars, then we can combine apparently diverse criteria. Brogden and Taylor (1950) have suggested the combination of varied criteria into a single measure of profits versus costs to the organization. Reminding management of the monetary importance of personnel decisions has great practical appeal. But, as Wallace and Weitz (1955) point out in their excellent discussion of criteria, converting all objectives to monetary terms is impractical. For one thing, neither indirect labor costs nor satisfaction of employees can be so converted; public relations goals such as "good will" can similarly not be translated readily into dollars and cents.

A third basis for combining is judgmental, either by direct estimations by policy makers of the relative importance of various criteria or by "capturing" operational policies by analysis of actual decisions made. (A good reference on policy capturing is Slovic & Lichtenstein, 1971.) Either of these has the advantage of taking the decision concerning weighting out of the hands of psychologists and putting it in the hands of the people who are paid to make policy. (This is not to say that psychologists should not help to serve as agents for change in policy, but only that they should recognize their consultative and, perhaps, persuasive roles.) Fiske (1951) argues that explicit policy and empirical research can substitute for judgment in development of criteria. We can only wish him good luck.

## ARGUMENTS AGAINST THE SINGLE CRITERION

Establishment of a single criterion appears to many as a hopeless quest. There are two reasons, one logical and one empirical. Logically, all components of a single criterion measure should be expected to represent different aspects of a unitary concept or different ways of measuring the same

characteristic. Combining such variables as production and absences seems like adding olives and toothpicks to obtain the alcoholic content of drinks. Especially clear pleas for multiple criteria have been made by Otis (1940), Toops (1944), Guion (1961), Weitz (1961), Dunnette (1963), Biesheuvel (1965), Guion (1965), Wallace (1965), Ronan and Prien (1966), and Roach and Wherry (1970). They argue that success is not unitary, for different jobs for the same person, for different persons on the same job, or for different aspects of the same job for the same person. Ghiselli (1960) points out that two persons may achieve equivalent total performance with quite different patterns of behavior, and hence, logically, evaluation of either treatments or individual differences should be made on the basis of different measures.

The second reason for rejecting the single criterion concerns the obtained interrelationships among criterion measures and predictive measures in empirical studies. An overwhelming majority of studies involving statistical analyses of sets of criterion measures finds that these analyses rarely yield a single general factor. In other words, several criterion measures are necessary to account for the variance in a criterion correlation matrix. To mention only a few studies, Ewart, Seashore, and Tiffin (1941), Rush (1953), Grant (1955), Stark (1959), Seashore, Indik, and Georgopoulos (1960), Forehand (1963), Ronan (1963), Schultz and Siegel (1964), Wiley (1964), Siegel and Pfeiffer (1965), and Kirchner (1966) all reported multidimensionality of criteria. This convergence of evidence should partly answer the arguments of Marks (1967), who argued that unreliability contributes to the "apparent" complexity of criteria. We shall see further evidence of multidimensionality as we survey the empirical results.

The argument has progressed to a consideration of how many criteria should be utilized. This question will prove to be an empirical one in each case. When a single composite will cover most of the variance in a set of criteria, then that composite should be used to represent them, and other criteria sought. When several dimensions are involved, several sets of criteria or composites will be required.

## CLASSIFICATION OF CRITERIA

We can classify criteria in a three-dimensional framework, as illustrated in Figure 1, which represents a cutaway diagram of kinds of criteria. The first dimension is the time span covered. The second is the degree of specificity of the criterion (as related to our discussion of multidimensionality above). The third is the closeness of the relationship to criteria based on the goals of individuals, organizations, or society.

### Time Span Covered

Criterion measures can be obtained either very soon after actual on-the-job behavior has occurred or many years afterwards. We may have, for example, a count of the number of times a worker stops his work for a rest break (Smith & Lem, 1955), which is very short-term, or a rating of success based on life history data including salary (Bingham & Davis, 1924), which is long-term. What is apparently the same criterion measure can involve different behaviors and abilities at different periods of time. It has been long known (Kornhauser, 1923; Blankenship & Taylor, 1938; McGehee, 1948; Smith & Gold, 1956) that performance early in the learning period does not necessarily correlate highly with later performance. More recent studies substantiate those findings (Ghiselli & Haire, 1960; Bass, 1962; Prien, 1966; MacKinney, 1967). These changes may represent changes in organizational demands with increased time on the job (Prien, 1966; Seashore & Yuchtman, 1967). They also may reflect a shift in the abilities being used. Factor structure shifts (Fleishman & Fruchter, 1960) and the correlation with predictive tests changes (Ghiselli & Haire, 1960) with time. All this leads to a plea for longitudinal studies (e.g., Ghiselli, 1956; Guion, 1967) and consideration of the dynamic nature of criteria. Too often, mere convenience, rather than the relevance to long-term performance, has dictated the use of measures.



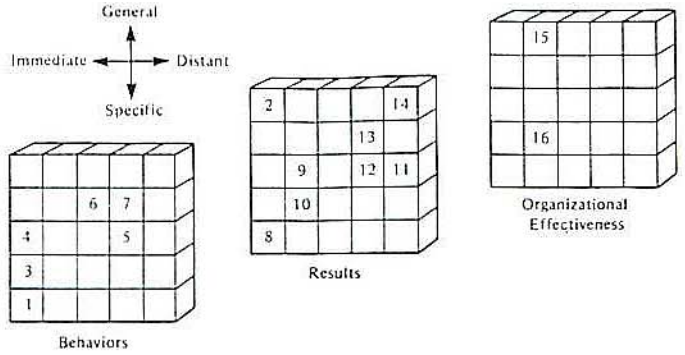


Figure 1. Dimensions of criteria.

The time span dimension has implications for the prediction of criterion measures. The time span of a manipulation or a predictor should be matched to the time span of the criterion measure. Thus, changes in the short-run situation, such as a bonus for attendance, could be expected to be reflected only in such short-term behaviors as absences and tardiness, and not necessarily in long-term job satisfaction. The latter might be more affected by policy toward promotions, which is long-term. There should be a match in time span. And predictors should be measuring those characteristics of individuals which will be relevant at the particular point in time in which criterion measures will be gathered.

### Specificity

Criteria vary also in their specificity-generality. Some may refer to very specific aspects of behavior (or effectiveness) on the job, while others give a summary estimate. Here the literature on multidimensionality of criteria is relevant (see above), although multidimensionality occurs across all of our classifications of criteria.

Regardless of the time span, criteria can differ in the specificity with which they refer to descriptions of performance versus global estimates. At one extreme we have a rater check whether a certain behavior has been observed (Flanagan, 1954) and on the other we have results combined into the composite dollar criterion (Brogden & Taylor, 1950). Another pair of examples is the tabulation of absence in a department (Stark, 1959) versus a composite of indices (Kurtz, 1937). Again there are implications for prediction and manipulation. Change in a single variable cannot be expected to have much effect on a general criterion. Since there are multiple factors causing change in general performance, the effect of any one will probably be small. Conversely, a test of general interest in the job cannot be expected to predict very precisely the specific criterion of absences. If possible, the relative degree of specificity in a manipulation or predictor should be matched to the specificity of the criterion measure.

### Closeness to Organizational Goals

Most important in classifying criteria is the dimension that concerns the closeness of

the required decisions in relation to organizational and societal goals. Criteria range from the description of actual behavior, through evaluation of results, to estimates of the effects upon the organization and society. Ideally, of course, it would be these last estimates which we would hold in mind in doing any research on policy making. Organizational goals such as economic stability, growth, and flexibility, and societal goals such as contribution toward individual well-being and growth, economic and social vitality of the community, and general productivity are the kinds of goals toward which our efforts are directed. But unfortunately, these are not always the dependent variables in which investigators are interested.

One reason is that such goals may appear to be inaccessible to the investigator. Actually, criteria and their measures can be developed for these goals. The practical difficulty is that investigators in general do not have the financial support for these types of investigations. Lacking this support, organizations and society are deprived of long-term projects and their evaluation in favor of short-term projects which may "pay off" sooner.

Another reason is that a particular manipulation or prediction is frequently much removed from organizational goals, can be expected to have only a minute effect on the total picture, and correlation with a measure of organizational effectiveness would probably be small, even though it might prove statistically significant. Moreover, a global or molar organizational evaluation would not permit diagnosis of the extents of achievement of more specific goals.

This dimension involves, first, the combination of specific human behaviors into generalizations about results (such as ratings or summary personnel statistics; cf., Guion, 1961), and second, the combination of a number of these generalizations to evaluate their impact on organizations, individuals, or society. Both steps involve problems, many of which will be discussed more fully below.

Here let us note briefly that the making of a generalization concerning results from specific behaviors involves a consideration of how they are to be combined. Whitlock (1963) indicates that there is a psychophysical law relating single behavioral elements to overall judgments. The log of the ratio of observations of effective performance to observations of ineffective performance is linearly related to the log of summary evaluative ratings, indicating that observations actually enter equally (without differential weighting) into a summary evaluation. Whether this is the manner in which observations *should* be combined is another matter. Perhaps a regression equation can do a better job than human judgment.

Ghiselli (1956) points out that there is individuality in the way components should be put together to obtain an overall estimate of results. Each individual can achieve results in his own manner, using different job behaviors and different abilities. Guion (1967) also points out that the dimensionality of criteria may be peculiar to each individual.

The step from results to organizational effectiveness is also large. If results which are not really important to organizational goals are weighted in making a judgment, then the final criterion is contaminated. If important aspects are omitted or not weighted enough, the criterion is deficient (Ghiselli & Brown, 1955). The evaluation of contamination or deficiency is judgmental. In my opinion, this judgment has to be made by management; this sort of decision is what management is paid for. But the psychologist can help in several ways.

He can ask directly for a judgment of importance. This judgment can be obtained either before or after the results of any comparison are in. After results have been obtained one can at least take into account whether certain aspects of the criterion are predictable or not (Kurtz, 1937). Weighting by importance has to be approached with caution, however (Ewen, 1967), since it, like other methods of weighting, may add nothing to validity. Rating of importance, however, is dependent upon the presence of the factor being rated (Smith & Landy, 1969). The psychologist can apply cost accounting principles to achieve a composite, as suggested by Brogden and Taylor (1950) and Likert and Seashore (1963). The difficulty here is that some aspects of organizational effectiveness are much easier to put into monetary terms than others, perhaps biasing the result in favor of the obvious and the short term. Or, he can apply statistical techniques to determine more clearly what are the bases of policy decisions. He can determine the weights actually given to various factors in making policy decisions on an intuitive basis, and systematize the weighting of those factors.

The large literature concerning policy capturing cannot be reviewed here. Only a small portion of the research has been concerned with industrial problems, largely for practical reasons. All the methods require a large number of stimuli to be judged by each judge or by a number of comparable judges, and hence the methods are typically applicable only in large homogeneous organizations. It is no wonder that such research has been conducted either in military settings or with manufactured stimuli (such as persons simulated in the form of personnel folders) rather than actual subjects to be judged.

Various procedures have been used to group judges according to similarity of their policies. JAN (Judgment Analysis, Christal, 1968; Naylor & Wherry, 1965) groups raters according to the degree of similarity or homogeneity of their multiple regression equations for predicting standing on a single criterion measure from the various cues they used in making their judgments (Williams, Harlow, Lindem, & Gab, 1970). Alternative ways of grouping by the the use of profiles of factors are PROF (Profile of Factors Method, Wherry & Naylor, 1966) and COPAN (Component Profile Analysis, Maguire & Glass, 1968), which use factors rather than item scores for greater reliability.

These methods show great promise for those situations in which sufficiently large groups of persons can be judged. An excellent review of such methodology is given by Slovic and Lichtenstein (1971).

In any case, we are trying to move from results to what Astin (1964) calls conceptual criteria, that is, "a verbal statement of important or socially relevant outcomes based on the more general purposes or aims of the sponsor" (Astin, 1964, p. 809).

## DIRECT OBSERVATIONS OF BEHAVIORS

Criteria are seldom actual records of behavior, unfortunately, and the behaviors "slab" is not heavily represented. We could, for example, go out on the production line armed with a stopwatch, time the elements of movement involved in the job, and summarize their variability by computing standard deviations. These could be used as indices of fatigue. The indices are on the behaviors slab, and quite specific and short term. The relationship to organizational objectives is remote. These measures belong in Cell 1 in Figure 1. McGehee and Owen (1940) timed actual rest pauses taken, as did Smith and Lem (1955), showing them to be responsive to short-term manipulations (scheduling of rest periods and size of work lot). These, too, are specific, belonging in Cell 1.

Flanagan (1949) established the critical incidents technique, in which specific job behaviors which are critical to satisfactory or unsatisfactory performance are elicited by interview of superiors, subordinates, and coworkers, and then translated into a check list of behaviors actually observed, so that these incidents can be summed to obtain an overall evaluation. (The end results belong on the results "slab," short-term, and gen-

eral; see Cell 2.) The technique has proved especially useful with personnel on whom a large number of incidents can be observed, as in the military. It would be difficult to adapt to the observation of managerial personnel, for example, for whom no two incidents may be alike. It has led to further attempts emphasizing direct observation of behavior.

Critical incidents have been used with favorable results by Kirchner and Dunnette (1957), among others. Kay (1959, p. 270) on the other hand found that "judges throughout actually were incapable of assessing the degree of likelihood of effective, average, and ineffective foremen doing that which was described in the critical incident." The problem seems to be that of opportunity to observe, rather than the format itself. We shall have more to say on that point later.

Another example of criterion measurement on the behavior slab is the interesting technique of Whitlock, Clouse, and Spencer (1963) in which observers tallied accident behaviors, or unsafe performances (see Cell 3, Figure 1). They report fairly high reliability for accident data but a relatively low correlation with actual injuries.

One that is much older, however, is the anecdotal report form, which stems from clinical and medical observation—the free verbal report of incidents (a technique for which we can give no primary reference). It involves someone's recording every incident of interest in a person's behavior. It can be commended for its thoroughness and condemned for its impracticality as it relies upon the conscientiousness of reporting by busy individuals, and, what is probably worse, their literary style. We cannot expect dramatic reporting from the ordinary observer. These reports belong in Cell 4, as immediate and somewhat more general.

But it is these observations that have to take place if our criterion measures are to be soundly based on fact. The trick is to obtain records of actual job behaviors before the process of selective recall distorts the impressions. Whitlock's (1963) psychophysical law indicates that favorable and unfavorable impressions acquire unitary weights in combination, which indicates the need for a more detailed basis of recording.

One approach is to ask observers to record at least sketchy notes on their observations on the job, not expecting them to write biographical portraits of the people they observe but merely to note a date and some reminder of the incident with some generalizations (see Cell 4, Smith & Kendall, 1963). (The job anecdote file suggested by Guion, 1965, is similar.) This approach by no means solves the problem of incomplete observations, since observation is usually squeezed in between many more urgent duties. It is at this point that many rating systems fail; they cannot enforce observation.

Another approach is to place a special observer into the situation with no duty except to observe and record. The observer can be either a participant behaving as if he were no different from any other worker [a good example was Owen's (McGehee & Owen, 1940) participation as a clerk in an office in which changes in rest pauses were introduced] or an obvious outsider, as in the work curve studies (Roethlisberger & Dickson, 1938; Rothe, 1946a,b, 1947, 1951; Rothe & Nye, 1958, 1959, 1961; Smith, 1953; see Cell 1). The use of an observer greatly improves the quality of the obtained data, but is costly and may disrupt the customary activities of the people being observed.

Another approach is to take the person being rated or observed off the job into a special situation. One type is the simulation of the job, used most conspicuously for the evaluation of the readiness of astronauts for space flights. It has been used successfully for evaluating proficiency in maintenance checking (Besnard & Briggs, 1967) in which there was no difference in errors between the simulator and the operational equipment groups. It has been used broadly in dangerous or highly critical work (see Cell 5).



Standard flights in pilot evaluation and a check list of items descriptive of a pilot's performance were evaluated in a summary by Viteles (1945), and have been used systematically more recently.

For executives, whose jobs are even more variable, the simulation has been largely on the basis of the In-Basket Test (Frederiksen, Saunders, & Ward, 1957) and assessment centers (see Chapter 20 of this *Handbook*). The In-Basket Test presents to the executive or manager a sample of incoming mail and memos, and records how he sorts, prepares for action, gets work accomplished, and seeks guidance from others. A complicated system, or rather several systems, of scoring are produced. Reliabilities vary from zero to high depending on scoring category. But the technique shows significant (if low) relationships to ratings from higher management (Meyer, 1970). It is a promising type of criterion at the managerial level. (See Cell 6.)

The assessment centers of large companies serve the same function and also take the manager or executive away from the job for observation. For example, the Sohio Assessment Program (Finkle & Jones, 1970) uses a job sample of managerial activities as well as peer ratings in the evaluation of progress and performance of managers, without any reliance on ratings by other managers. The program is aimed at assessing discrete aspects of performance, assessed from tests and situational exercises. Thomson (1970) reports much greater reliabilities at the assessment center than in the working situation; the actual supervisors were more lenient and demonstrated more halo (see below) than did the observers at the center. (Perhaps the observers were concentrating more upon observation.)

Grant and Bray (1966) report the extensive assessment center activities at American Telephone and Telegraph in which interviews, projective techniques, and situational tests were used with significant success to predict managerial effectiveness. Again, observation of behavior is intensive, intentional, and predicts later performance. Assessment centers are somewhat distant and general. (See Cell 7.)

We cannot emphasize too strongly that observation of actual behavior on the job or, if necessary, off the job is the core of establishment of a successful criterion. Without careful observation, we cannot make valid ratings or evaluate the meaning of so-called objective or "hard" criteria discussed below.

## EXAMINATION OF RESULTS

The results "slab" contains two sets of criteria—the "hard" criteria obtained from organizational records such as absences and turnover, and the "soft" criteria obtained from ratings. The first maintains the appearance of objectivity; the second is frankly judgmental.

### Hard Criteria

Company records furnish data for evaluation of performance, both specific and general and long- and short-term. They represent material available from payroll, insurance, and personnel records, without explicit use of ratings or other evaluation, although, as we shall see, evaluation has entered into every figure in the books.

#### *Tardiness*

A short-term specific criterion is tardiness. Its short-term characteristics are emphasized by one of the very few studies using tardiness as a criterion (Mueser, 1953). It was not rain, but bright sunny days that were related to tardiness. It belongs in the lower left corner of the second "slab" in Figure 1 (see Cell 8).

#### *Absences*

Absences can be measured in a number of different ways. The base may differ as in

percentages of scheduled working hours absent versus number of occasions absent (regardless of the number of days involved in an occasion). They may be broken down into absences attributable to different causes, as illness, personal, excused or unexcused, despite the difficulty in ascertaining the actual cause. In any case, indices of absences behave differently as criterion measures both among themselves and as compared with other criterion measures, and with predictive measures. Kerr, Koppelmeir, and Sullivan (1951) found, for example, a correlation of −.44 between unexcused absences and job satisfaction, while total absences correlated .51 with job satisfaction. (There was no control for job level.) Metzner and Mann (1953) also found frequency of absence superior as a criterion to the actual days lost. Huse and Taylor (1962) report that the total absence frequency is the most reliable absence measure.

Situational factors may greatly affect this index, as is indeed true for all the so-called "hard" criteria. Behrend (1953) points out that absence rate is affected by the labor market conditions at the time. Stark (1959) points out that absences may be a function of factors beyond the control of the manager, and absences are used not only as a criterion of individual performance but also as a criterion of effectiveness of foremen. The relationship of absences to personal history and organizational characteristics is different in larger than in smaller units, and for blue- and white-collar workers (Baumgartel & Sobol, 1959). Argyle, Gardner, and Cioffi (1958) note that absenteeism was not related to either turnover or productivity (using departments as the units of analysis), which agrees with Ronan's (1963) finding that the factor structure and weightings of various criteria shift from plant to plant, and with the situational variability reported by both Lyons (1972) and Porter and Steers (1973). Clearly, absences should not be blindly adopted as criteria. They should be located as fairly short-term, medium specific results (see Cell 9), and should not be expected to relate closely with more general, long-term organizational goals.

### Accidents

This notoriously unreliable set of measures (Ghiselli & Brown, 1955, p. 344, for example) is nevertheless important to some organizational goals, including humanitarian and economic ones, to which fingers can be easily pointed. The problem is that most accidents are, by definition, beyond any person's immediate control. The best solution, in all probability, is to return to the behaviors "slab" and observe accident behaviors (Whitlock, Clouse, & Spencer, 1963). Barring this more careful inspection of behavior, statistics need to be compiled on the actual record, which is not a very satisfactory process if one is interested in individual accident performance. Accident statistics based on group data are more reliable. Daniels and Edgerton (1954) validated ratings by superiors against the percent of damaged vehicles in motor units, and found a significant relationship. But Ronan (1963) again finds no consistent factor structure for his injury index (visits per year) and lost time accidents in his analysis of multiple criteria.

The problem of the base for accident figures is worse than that for absence figures. One can compute accidents per hours worked, pieces produced, miles driven or flown, or trips taken, giving entirely different results. Like any other errors, including mistakes in executive decisions, more errors will be made with more opportunity for error, and hence the rate should be taken into account. On the other hand, it seems reasonable to penalize a person for moving too rapidly and taking unwarranted risks of error.

The decision to use one base or another should be a logical one having to do with what treatments, conditions, or measures are to be related to the criterion. Accidents are relatively immediate, specific results. (See Cell 10.)



### Tenure or Turnover

Length of service has been used as a criterion in many test validation studies, and has proved to be predictable. A complete review of the literature up to 1964, including the studies of predictive validity, is given by Schuh (1967). Turnover, sometimes defined as the number of terminations divided by the average number in the working force and sometimes as the average number of terminations and accessions divided by the average number in the working force, is used to assess the effectiveness of groups, or of supervisors of groups. In any case the decision of an employee to terminate is a long-term decision, but moderately specific, and belongs at the right middle of the results "slab" (see Cell 11). It should not be expected to respond to minor changes, such as shifting the schedule of rest pauses.

Turnover is related to the alternative job openings that are available (Behrend, 1953; Tiffin & Phelan, 1953; Stark, 1959) and hence may reflect factors beyond the control of management. It is apparently not related to absences but is related to productivity (Argyle, Gardner, & Cioffi, 1958). It will continue to be popular and important as a criterion because of its obvious relationship to costs and returns, and hence, to organizational goals.

### Sales

For evaluating performance of sales personnel, actual dollar sales are an obvious choice as a criterion. But it, too, has pitfalls. For example, Rush (1953) found four factors in fifteen scales measuring sales knowledge and performance in different ways, three of these involving different measures of sales. It makes a difference, for example, whether we measure percentage of quota achieved or average monthly volume. Taylor, Schneider, and Symons (1953) found similarly that basic salary for salesmen was a more predictable criterion than bonus earnings (in a system supposedly adjusting for differences in ability). When we go to studies of differences in groups, we find fairly good reliability of measures of group turnover and sales productivity per man (Weitz & Nuckols, 1953). State by state, clusters of nineteen measures showed five factors, including the absolute size of the state. This factor illustrates an important problem in the use of sales as a criterion: It should be adjusted in some way for the potential of the sales territory. Ideally, we would have actual norms for each territory, so that performance could be compared with the norms, but accumulation of such data is difficult, and ratings of difficulty of territory have to be substituted. This criterion would appear to be a fairly long-term, fairly general result. (See Cell 11.)

### Production

Direct measures of output would seem to be closest to organizational goals and most desirable to use as criteria. They are short term and moderately general. (See Cell 9.) They are very effective in the ideal (for psychologists) situation in which there is only one job in the entire population to be examined, as in the handkerchief factory studied by Smith and Gold (1956) in which half the employees ran the handkerchiefs through the machines the long way and the other half ran them crossways with the same average production for each group. But that is not at all typical in the working world. Often, in order to enhance sample size, similar but different jobs are combined and must, therefore, be put on a comparable basis. This equating requires that for each job we estimate what is "normal" production, and express productivity as a ratio to normal or standard. Herein lies the rub.

Time study of the jobs must precede the setting of standard rates of production, and this study must include some rating of the effort and skill of the person who is observed and timed. (For a clear description see Krick, 1962.) This rating is rife with errors of rating (see, for example, Argyle,

Gardner, & Ciotti, 1958; Lifson, 1953; Ryan, 1947). These errors are perhaps one reason that records of production have not proven to be as useful or as popular for criterion purposes as was once hoped (Schultz & Siegel, 1961).

A related criterion is the length and shape of learning curves. Lefkowitz (1970) showed the effect of training on productivity of sewing machine operators, as did Smith and Taylor (1956). In both studies the progress of learning was not linear. Thus, a criterion measure should be taken in relation to the actual learning curve rather than a linear rate of progression. It is also probable that learning curves should be evaluated for each element of a complex skill, since certain elements respond more to practice than others do (Barnes, 1963). Rates of improvement in test performance also may occur at different rates for various elements making up the performance (e.g., Salvendy, Seymour, & Corlett, 1970). Again, the shapes of acquisition curves are discerned only with difficulty, and the criterion should be developed only with the help and judgment of a number of careful judges. They are relatively long-term and moderately general results (see Cell 12).

### *Job Level and Promotions*

The extent to which an individual reaches a high or higher job level has been used as a criterion, a fairly long-term and global or general one, on the results "slab" (see Cell 13). The extent to which promotions are a valid criterion is limited by the fact that many factors other than performance may affect promotions, such as political expediency, organizational structure, and labor market conditions. Nevertheless, promotions represent a chips-on-the-board decision concerning the value of a person to the organization. Job level should be evaluated in terms of some standard job evaluation scheme. However, actual promotions are frequently not based on performance evaluation, but rather word-of-mouth and other informal evaluations, and hence, may reflect many situational factors (Campbell, Dunnette, Lawler, & Weick, 1970). Nevertheless, job level has been used by several investigators including Henry (1948) and Bentz (1968) with success. It should be corrected for years in service as we shall see below.

### *Salary*

The same rationale—that actual operational decisions have been based on performance—makes salary level an appealing criterion, particularly at the managerial and professional levels. Bingham and Davis (1924) and Gifford (1928) used dollars as a criterion. The recognition of the importance of years on the job appeared explicitly somewhat later. A significant, if inadequately statistically documented book, is that of Jaques (1961), concerning the normal progression curves for equitable payment. A follow-up in the United States, accompanied by means and variances, is certainly warranted. Jaques examined, for a number of employees, their salary gains and computed lines of best fit for persons starting at a given initial level. Relationships between obtained curves and extrapolated curves are certainly impressive. Years seem to be related curvilinearly to gains in salary, but, more importantly, the relationship seems to be predictable. It takes a major change in job level to break the steady normal progression of salary with age. (Don't take a job below your proper asking price; you won't make it up later.)

This steady increase with age has led to corrections of job or promotional level according to years of tenure. Hulin (1962) used a number of corrections, the most sophisticated of which was probably obtaining the difference between salary increase and the salary increase predicted on the basis of tenure.

Even corrected, this criterion runs the risk of contamination, since many factors besides individual merit may influence salary. For example, internal politics, or



scarcity of personnel in a special field may affect the situation. Nonetheless, it represents a long-term, global result (see Cell 14).

## Soft Criteria

As we have seen, the so-called hard criteria all involve some subjective components. Human judgment enters into every criterion from productivity to salary increases. Merit rating as well as evaluation of causes of accidents or absences involves a subjective evaluation. In this process, some common errors exist. Many rating procedures have been developed which attempt to reduce or eliminate these errors.

### *Common Errors*

LENIENCY AND SEVERITY. The first common error is that of leniency. Ratings tend to be bunched toward the favorable end of the rating scales. The average person is rated as above average, making for a displacement of the mean, and skewness. The reasons are multiple. In the first place, there is real selection of persons to be rated. The worst have actually been fired or transferred. Moreover, the rater is usually in the position of judging his own competence along with that of the person being rated; if a superior rates his subordinates as incompetent, it reflects upon his own competence as a supervisor. These factors, together with normal human kindness, make for a bunching at the favorable end of the rating scales (Thorndike, 1949; Bass, 1956; Sharon & Bartlett, 1969).

The opposite can also occur: One way to appear good is to devalue the people around you, and this mechanism can affect ratings—the error of severity. This error can be detected only by a low mean rating and positive skewness. Again, this error is very hard to detect because what is really needed is a number of ratings of different groups.

SEQUENTIAL EFFECTS. The judgment of an item on a rating scale may be affected by the items which precede it, either more or less favorably. This error can be controlled only by randomizing orders of presentation.

DISTRIBUTION ERRORS. These represent deviations from the expected more or less normal distribution curve. The errors usually indicate failure to discriminate. When persons or products are being rated, the ratings tend to pile up in the middle of the distribution. (This error may be compounded with the errors of severity or leniency to give a leptokurtic and skewed distribution.) When items describing people are being rated, raters tend to pile up *items* at the *ends* of the scales, avoiding the middle (Cliff, 1959; Rotter & Tinkleman, 1970). It is difficult to write neutral items, as has been reported by attitude scalers, although Obradović (1970) seems to have managed to locate items of approximately neutral attractiveness. The trick seems to be to write double-barreled statements or very general ones. Anchoring statements on the scale on which items are to be rated will shift the distribution—a positive anchor making for a shift from neutral to negative, and vice versa. In any case, the distributions of ratings should be checked before they are used.

INTERCORRELATIONAL ERRORS. The first such error is the all-pervasive halo effect, which means that rating on one characteristic spills over to affect ratings on other characteristics, resulting in high intercorrelations among ratings for supposedly different characteristics or behaviors. Halo can be either favorable or unfavorable—it merely represents the failure of the rater to differentiate. The prevalence of the error has been frequently reported. Two examples will suffice. Turner (1960) factor analyzed twenty different measures of criterion performance of foremen, and found two factors, one involving all the ratings (halo), and the other involving employee relationships, including nonrating measures. Vielhaber and Gottheil (1965) found, amusingly, a correlation of .31 between ratings based only on name and home

address and ratings by peers and superiors after fourteen weeks at West Point.

Contributing to halo are the effects of contrast and similarity. Some raters rate in relationship to their own self-ratings—either contrasting others to themselves or assuming similarity to themselves. Training seems to be the solution.

The logical error is more difficult to handle. It is difficult to tell when it is an error and when it is a legitimate inference. The individual rater has a pattern of correlations which he assumes among traits in others, and which remains when the effect of halo is removed through partial correlation (Koltuv, 1962). Without such a set of assumptions, hardly any rating would be possible; the problem is to eliminate false generalizations or at least to systematize the assumptions held. There is considerable agreement concerning the structure of personality, both for complete strangers and for persons well known to each other (Passini & Norman, 1966, 1969). This fact gives some hope for retaining the valid intercorrelations and eliminating the idiosyncratic ones. Psychologists have been misled, in my opinion, by the hope that traits and factors on jobs would prove to be orthogonal, when in reality they are intercorrelated to greater or lesser degrees. This intercorrelation does not mean that they cannot be assessed separately; it means only that the final results will be interrelated.

### *Types of Rating Scales*

Rating scales can be classified along each of the dimensions we have proposed for criteria, and scales can be constructed to fit any of the cells we have discussed. They can be directed toward very short or very long time spans. They can be very specific, or quite global covering only an overall estimate of performance. And they can be directed toward behavior or toward organizational goals. But they need to be classified further.

Most of these classifications have to do with format which may have relatively little effect on the resulting evaluation. Blumberg, DeSoto, and Kuethe (1966) found remarkably small effect of rating scale format, although raters, ratees, traits, and some of their interactions were significant. Stockford and Bissell (1967), on the other hand, found "a marked influence on the value of ratings" when descriptive scales were compared with the less effective evaluative scales. This finding agrees with that of Yuzuk (1961). Madden and Bourdon (1964) also found differences in results from different formats, with no general principle emerging. Hence, we must consider format seriously.

Direct estimation. Direct estimation of evaluative level is made using formats that ask directly the question, how good is the ratee? The answer may be recorded on a graphic scale typically running from left to right with a few verbal anchors such as excellent, good, average, poor, and unsatisfactory placed beneath the scale, although in its pure form it has anchors only at the extremes. Some more specific scales have been constructed vertically, with the good end at the top (Champney, 1941; Smith & Kendall, 1963), which has the advantage of permitting more detailed verbal anchors to be inserted along the scale. In any case, the method is characterized by estimating psychological distance directly by measuring distance along the linear scale. Numeric scales evaluate by asking for a number to represent the level of performance, and are usually combined with some verbal anchoring system (Blumberg et al., 1966). Similarly, alphabetic scales use letters of the alphabet (which are characteristically later transformed into numbers). Some of the direct estimation scales use verbal anchors directly, and later transform these into numbers. A novel symbolic rating scale is the General Motors "Faces" scale (Kunin, 1955), which represents a series of faces from a scowling frown to a pleased grin. Checks made on this symbolic scale, too, are transformed into numbers.



Ranking. Ordering of persons can be achieved, without asking for direct estimations of distance along a scale, by some version of ranking. For example, persons to be evaluated may be simply ranked from best to poorest along any relevant dimension(s). These ranks may be treated under a number of assumptions (see Guilford, 1954, pp. 178–195) to give more direct estimations of distances between persons, and to permit the combination of rankings of different persons by more than one rater. Or every individual (ratee) may be paired with every other individual (ratee) to permit pair comparisons, and respondents' preferences may be treated (see Lawshe, Kephart, & McCormick, 1949; Guilford, 1954, pp. 154–176; Edwards, 1957; Torgerson, 1958) to yield psychological distance between the ratees. Though one of the best of the scaling methods psychometrically, this method has the disadvantage of requiring a large number of judgments per judge (all possible pairs of ratees) and considerable computation. Another method to achieve ordering is the use of forced distribution ratings—requiring the rater to put a fixed percentage of persons into each of several categories, on the assumption of a normal distribution. This assumption is so seldom warranted in the actual industrial situation where people have been discharged, promoted, and cajoled to increase productivity in relation to their previous performance that this method is not recommended. The method has also been shown to be affected by rater bias (Klores, 1966).

Test construction methods. These can be used to achieve a scale of cumulative points. This approach implies that norms will be used to achieve an estimate of distance between persons. The most common procedure is to use an unweighted check list in which adjectives or descriptive statements are merely checked as applying or not applying to the individual being rated. Each favorable response is given a positive weight and each unfavorable one typically a zero weight to give a total score. The old scale by Hartshorne and May (1929) is an example of an unweighted scale. So also is the very much more sophisticated critical incidents approach (Flanagan, 1954; Buel, 1960), which involves collecting examples of good and bad behaviors which are to be used as a check list for evaluation. The advantage of this approach is the emphasis on observation of behavior. The disadvantages are the difficulty in obtaining actual concrete behaviors that can be observed and noted. Except in the military situation, it has become necessary to use broad generalizations, losing the huge advantage of emphasis on observation.

The use of expected behaviors (Smith & Kendall, 1963) rather than actual observations at least permits observers to generalize from specific observations to other specific predictions of behavior so that actual behavior is incorporated in the rating procedure. That judges indicate the dimension to which each expected behavior pertains facilitates clarity of scale definition. The scaling of expectations permits one method of weighting behaviors.

There are several approaches to weighting of a scale. Some examples of weighted scales are those of Ferguson (1947), Knauft (1948), and Uhrbrock (1950, 1961). Ferguson (1947) scaled a number of statements about managers by a modified method of equal appearing intervals (Richardson & Kuder, 1933) to establish a weighted scale of success. This method gave a scale that proved useful in evaluating managers in a variety of situations with different raters reporting. Knauft (1948) simplified the procedure by eliminating the comparison with an external criterion measure and established alternate forms of scales for two jobs. The last big step in weighting of items was made by Uhrbrock (1961), who scaled 2,000 (presumably) all-purpose items. (His first scaling was concerned only with foremen, but this restriction is not made clear in his last report.) His efforts were aimed at forced-choice matching of items (see below), but his items, scaled by equal-appearing

intervals, could just as well be used for a weighted check list. On the assumption of an underlying overall dimension of goodness, he has furnished us with a pool of items for use in constructing a weighted check list of descriptions of behavior for almost any job—so long as we wish a global rating.

Any of these can be transformed into an unweighted scale, simply by using unitary weights. There is no strong evidence that weighting actually improves the psychometric properties of a scale. A common approach is to "score" a ratee according to the median scale value of the items chosen by the rater as being descriptive of his job performance. A potential problem in the use of the median has been pointed out by Jurgensen (1949). If an individual ratee has been given credit for a large number of positive items, his median score will actually be lower than that of a ratee who was endorsed for only a few very highly rated items. Jurgensen recommends instead a strictly algebraic weighted score.

Semantic differential scales are a type of weighted check list in which weightings are obtained at the same time as ratings. The individuals are described along seven-point bipolar scales covering a dimension of meaning (Osgood, Suci, & Tannenbaum, 1957). Performance is described according to locations on scales defined by pairs of bipolar adjectives (such as *good . . . bad*). The scales are not really equal-interval scales (Heise, 1969), and they are not *always* bipolar (Mordkoff, 1965). There seems to be no great advantage over simple unweighted check lists or direct numerical rating scales.

ITEMS AS SCALED STANDARDS. The psychologist may set up a series of items as scaled standards against which an individual may be judged. These items may be used as anchors, usually along a graphic rating scale. For a global rating scale, Uhrbrock's (1961) items seem ideally suited. I once attempted to scale adjectives by the method of paired comparisons so that they might be used as scaled anchors for an evaluative rating scale. A deficiency of the method became apparent: the adjectives were pre-screened by normal item analysis techniques for low dispersions of ratings; the items became unscalable by paired comparisons because there was insufficient overlap of judgments.

Forced choice is a special technique designed to reduce the effects of rating errors and faking (Sisson, 1948). The procedure involves the presentation of groups of items matched for general desirability but differentiated in terms of their predictive performance against some overall criterion (for formats, see Guilford, 1954, pp. 154–176). Forced choice ratings are perhaps more valid and less susceptible to faking (Taylor & Wherry, 1951; Izard & Rosenberg, 1958; Zavala, 1965; Scott, 1968) than more direct rating scales, but still subject to bias (Travers, 1951; Kay, 1959; Howe & Silverstein, 1960; Howe, 1960). Four positively worded statements proved a valid, reliable, bias-resistant and acceptable format (Berkshire & Highland, 1953). There is some convergent validity (Taylor, Schneider, & Clay, 1954; Prien & Kult, 1968). The use of items of neutral attractiveness proved valid (Obradović, 1970) against a forced distribution rating. The forced choice format requires pretesting items for desirability, and is based on the belief that raters must be deceived concerning what they are rating and the interpretation of the rating. It also gives a global summary rating of performance. It is opposed to the multiple-measured approach recommended in the present presentation.

The forced-choice format attempts to overcome the errors of leniency or severity by concealing from the rater the meaning of his ratings. This is done by constructing rating scales such that it is not immediately obvious how different responses are scored. This effort to conceal the scoring system has, in practice, often led to a contest between the administrator and the rater. A rater, intent on giving favorable ratings, can beat the system, however, merely by rating each person in the same way as a formerly rated and high-scoring person. Any attempt to interpret scores or to give feedback to the ratee is also made very difficult by the concealed scoring system inherent in the approach.

The matching of alternatives according to social desirability also poses a problem since judged desirability depends upon the situation and the judges. The whole field of attitude measurement depends upon these differences, which undercut forced-choice matchings.

Forced choice also introduces another systematic source of error. In a typical combination of four alternatives in forced choice, two are dead items, used only as decoys. These items also attract positive endorsements. For example, "honest" may be a nonscored (nondifferentiating between high and low ratees) item, although it may be legitimately endorsed about a very good candidate. This effect introduces error variance. I prefer a more direct and more specific method of evaluation.

Results from forced-choice ratings belong at various places on the cube, depending entirely on the items included.

Another technique is that of scaled expectations (Smith & Kendall, 1963; Maas, 1965; Kendall & Hilton, 1965; Dunnette, Campbell, & Hellervik, 1968; Fogli, Hulin, & Blood, 1971), which sets up a scale for each dimension to be rated on desirability of behaviors which the rater might expect a ratee to demonstrate. These scaled behaviors serve as anchors for raters who have agreed upon the desirability and the dimension which each item represents. The scales are used with raters similar to or preferably the same as those who construct the scales. The advantages of the procedure are use of the raters' own terms, the specificity of the behavior rated, the emphasis upon observation, the lack of ambiguity of meaning of the anchors, and high scale reliabilities (in the upper .90s). The disadvantage is the necessity for collaboration of a number of raters in constructing the scales —a disadvantage that may well be a psychological advantage due to the training achieved.



INDIVIDUALS AS SCALED STANDARDS. An early example was the man-to-man scale (Guilford, 1954, pp. 269–270). This scaling method used actual people as anchors against which employees were to be compared. Ideally, the same anchor men should be used in all departments doing the rating, permitting comparison from department to department. The difficulties encountered were primarily lack of knowledge of the same key men by different raters and inequality of units from man to man. Ross (1966) has solved these problems and proposed a modern man-to-man scale suitable for overall, global ratings of performance based on a two-pronged approach. Each rater in an organization ranks not only his own subordinates, but also non-subordinates (reference persons) in other departments with whom he has had recent contact who are comparable to benchmark persons in his own department. Moreover, all evaluations of benchmarks are transformed to $V$ values, a type of standard score based on the rank of the individual within the reference group, and his overall standard score (transformed from ranks within his own department) for general performance, job level, and educational level. The procedure permits cross-referencing of ratings by different raters with different groups of subordinates, and takes into account the anchoring of individual rating scales. It seems to be a practical solution to the criterion problem when a single global criterion is wanted, although it would seem to be unwieldy for the construction of separate scales for different characteristics.

CONSTRUCTING SCALES. For most criterion problems, scales have to be constructed especially for the particular research or administrative purpose involved. Overall evaluation is suitable for discharge, promotion, and similar decisions. Most research, counseling, training, or transfer decisions require

separate scales for separate dimensions of performance. Several investigators have recommended that the scales should be descriptive rather than evaluative (e.g., Flanagan, 1954; Smith & Kendall, 1963). Stockford and Bissell (1967) report that descriptive scales are "more reliable, less influenced by bias, and show less deviation in leniency and severity than is characteristic of ratings on 'subjective' or *evaluative* scales." Scales should be defined briefly and so also should levels within each scale (Madden, 1964). There is disagreement about whether it is worthwhile to scale the anchors more finely than favorable-unfavorable (e.g., Bass, 1956), although the idea of scaling has proved to be beguiling (Uhrbrock, 1950, 1961; Smith & Kendall, 1963). Actually the entire attitude-scaling literature is relevant to measurement of attitudes toward employees (cf., Edwards, 1957; Torgerson, 1958; Fishbein, 1967).

There are two psychometric problems that arise in anchoring scales: locating defining statements on the correct scale and at the right place on that scale. An attempt to solve these problems is retranslation of expectations. Briefly, the persons who are to rate indicate and define the dimensions along which descriptions are to be made. They then write examples of behavior to be expected of persons along each scale. Other raters independently allocate these examples to the dimensions to see if there is agreement as to their meaning. Independently, they are assigned numbers to indicate position along the relevant scale. Examples (and scales) about which there is low agreement are eliminated (Smith & Kendall, 1963; Fogli, Hulin, & Blood, 1971). The procedure leads to examples with high scale reliabilities.

### *Training of Raters*

The success of these anchoring procedures depends in my opinion on the training of the raters. Attempts to use the scales by persons who had not been trained in rating resulted in relatively low success (Hakel, 1966).

Training should include instruction on the principles of rating, participation in selection of items and follow-up. Such training "raises the reliability of, and reduces the effects of bias on, the merit ratings that are given their subordinates" (Stockford & Bissell, 1967). In the Smith and Kendall (1963) study, extensive conferences were held covering rating errors, observation, and implications of observed behavior for future behavior. Raters participated—and that is important—in defining the areas to be evaluated, in defining the levels of each area, in writing examples of behavior, in allocating examples to areas, and in indicating the scale value of each example. A crucial aspect is the observation and recording of relevant behaviors. Another important aspect is the fact that conferences on evaluation seem a natural forum for participative discussion of supervisory practices.

### *Who Should Observe and Evaluate?*

No amount of training can improve a rater if that rater has had no opportunity to observe the ratee's behavior. This fact makes the choice of the rater crucial. The choice is usually made on the basis of expediency rather than potential accuracy of rating. Yet the rater is more important than the technique (Bayroff, Haggerty, & Rundquist, 1954).

Qualification of raters. There are some data on qualifications of good raters. The obvious takes place: superior intelligence and effectiveness are associated with less biased and more reliable ratings (e.g., Schneider & Bayroff, 1953; Stockford & Bissell, 1967). Some people are consistently better in terms of validity of ratings than others (Wiley & Jenkins, 1964). Familiarity of rater with ratee is also a key factor in quality of rating (Besco & Lawshe, 1959). The relation of initiating structure to consideration is related to variability of ratings, emphasis on production, overall level of rating, and absence of leniency (Klores, 1966).



More effective managers valued initiative, persistence, broad knowledge, innovation, and planning as contrasted with less effective supervisors, who emphasized cooperation, company loyalty, teamwork, accepting suggestions, and tact and consideration (Kirchner & Reisberg, 1962). Stockford and Bissell (1967) found that in their study the ratings reflected "primarily the personal-social relationships between supervisor and subordinate rather than the output of the subordinate in question."

Superiors. The answer to the question of who should rate has usually been the immediate supervisor, and possibly a more remote one. All of the previously cited literature concerns such ratings. They have the advantages of face validity, acceptability, and availability, although it is frequently difficult to find more than one rater familiar enough with the ratee to achieve an independent estimate of reliability. There is frequent reason to question the actual opportunity by the supervisor to observe behavior on the job. There are alternatives.

Peers. The most popular alternative is peer rating. Peer (or "buddy") ratings take two principal forms—the nomination of peers as best or poorest on some dimension or ranking from best to poorest (Hollander, 1954). The earliest applications were to military leadership and were concerned with the *prediction* of success in Officers' Candidate School and combat performance (with notable success). The early evidence is well summarized by Hollander (1954). The ratings do not reflect merely popularity (Wherry & Fryer, 1949; Hollander, 1956, 1965). Nevertheless, Doll and Longo (1962) corrected peer ratings for perceived "anti-social" characteristics of ratees to increase predictive validity. Peer ratings are valid even when administered with an administrative set (Hollander, 1957). Peer ratings are more stable over time than supervisors' (sergeants') ratings (Gordon & Medlund, 1965). French (1954) developed a single reliable criterion based on a combination of ratings by officers, upperclassmen, and classmates.

Civilian applications are fewer. Again, predictive validity was established against promotion (Roadman, 1964). Tucker, Cline, and Schmitt (1967) found no convergent validity between peer ratings and superior ratings. Peer ratings seem to be tapping a different source and/or type of information even though French (1954) did see fit to develop a combined criterion from peer and superior ratings. Peer ratings should probably be examined as separate criteria particularly for leadership (Wherry & Fryer, 1949). Additional industrial applications are needed, especially in identification of leadership potential. The threat to management of allowing nominations from the rank and file is undoubtedly responsible for hesitancy in the use of peer ratings. They should, nevertheless, be considered seriously. Peers, after all, have the opportunity to observe. Borman (1974) showed, in fact, that peers and superiors not only provide different perceptions of job performance but that they actually evaluate different aspects of performance in very different ways. Moreover, superiors' ratings of subordinates suffered reduced reliability when the superiors used scales which had been developed by subordinates. Similarly, the peer ratings provided by subordinates of each other suffered from low reliability when they were based on the scales developed by superiors.

Subordinates. Ratings by subordinates do not agree with superiors' ratings (Fleishman, Harris, & Burtt, 1955; Rambo, 1958; Besco & Lawshe, 1959; Tucker et al., 1967) although they identify some aspects of promotability (Mann & Dent, 1954). Besco and Lawshe (1959) compared ratings of leadership (Rambo, 1958) by employees and general foremen with ratings by higher managers concerning departmental effectiveness. Superiors' ratings were related only to consideration. Superiors' and subordinates' ratings are tapping different dimensions of performance. Perhaps more information can be obtained from the classroom situation,

where the resistance to subordinates' ratings is probably less.

SELF. Self-ratings have been even less used—except in the enormous job satisfaction literature. In performance evaluation, they have been avoided because of the obvious possibility of bias. Kirchner (1966) found a significant (if low) relationship between superior and self-ratings. Thornton (1968) found a bias in favor of self, and a relationship of that bias to lack of promotability, despite a positive correlation on some characteristics between superior and self-ratings. Lawler (1967) on the other hand found disagreement between self-ratings and both peer and supervisors' ratings. The use of self-ratings in performance appraisal interviews is emphasized by Bassett and Meyer (1968). They used self-ratings as a basis of an interview with the ratee, and report less defensiveness and fewer complaints with the procedure. It is in their sort of counseling setting that self-appraisals seem most promising.

JOB SATISFACTION. Job satisfaction is a goal in itself. It has been evaluated in numerous ways (see Chapter 30 in this *Handbook*) with differing results. We can refer again to Figure 1. Job satisfaction measures can vary, particularly on the time span dimension. Measures taken in the framework of a short time span can be expected to relate to short-term behaviors such as tardiness or rest pauses, while measures with a long-term reference can be expected to relate to long-term behaviors such as terminations and job choice. Job satisfaction should serve as a criterion for evaluation of such treatments as leadership training, organizational structure, communications networks, and job enrichment. The self-evaluation seems to be the logical choice for such criteria.

ASSESSMENT OBSERVERS. Evaluation can be made by psychologists or management personnel in assessment centers, in which situational exercises may be engaged in, testing may take place, and group discussions may be held. It is difficult to determine whether the measures are criterion or predictive measures; they are sometimes one and sometimes the other. That they can have predictive validity against more remote criteria (about Cell 7 against Cell 13) has been demonstrated by Grant and Bray (1966). Predictive validity has been demonstrated also for salesmen (Bray & Campbell, 1968).

## EXAMINATION OF OUTCOME: IMPROVING INFERENCE

We now proceed to the outcomes or organizational effectiveness slab. Here we are concerned with decisions based on inferences from results. These inferences involve moving from either hard or soft criteria to decisions concerning organizational effectiveness. They usually involve combining several criteria.

### The Problem of Weighting

#### *Direct Weighting by Management*

Management makes the decisions concerning courses of action to be taken on the basis of evidence. But the choice of how to weight different aspects of evidence to form the decision is usually quite unsystematic. The psychologist can help by asking for formal judgments of the amount of weight to be placed on various aspects of results or evidence contributing to the decision. These judgments may be either objective or highly subjective. The mere formalization of policy helps, but is difficult to obtain.

#### *Policy Capturing*

Sometimes the psychologist can help by using statistical techniques to "capture" the weighting given by management to various items of information. We have discussed policy capturing previously, but here we are concerned with its application to organizational effectiveness. Here, artificial stimuli or simulated aspects of evidence believed to be part of the decision matrix can be presented to executives who will be asked to make hypothetical decisions based on their own views of which combinations may lead to greatest organizational effectiveness. This essentially is the strategy used by Borman and Dunnette (1974) when they sought judgments from Naval officers about the relative effectiveness of the personnel subsystems of 100 simulated ships. Then, the decision process can be made more systematic in at least two ways, based on the predictive equations summarizing the judges' evaluations: First, future decisions might be turned over to computers; and second, feedback could be given to the executives in training sessions focused on helping them to improve the consistency of their own judgments (Slovic & Lichtenstein, 1971). In these days when complex decisions related to such matters as air pollution and public opinion have to be balanced against immediate corporate profits, research in this broad area is sorely needed.



#### *The Dollar Criterion*

The use of immediate corporate profits as a criterion becomes less and less plausible. We need criteria that reflect the societal and long-term organizational goals as well as the economic ones. It is important, nonetheless, that the psychologist working in industry apply some cost accounting to his work—that is, if he prefers to retain his job during recessions. He is referred to Likert and Seashore (1963) and a good accounting text for his do-it-yourself kit. The dollar criterion, as Brogden and Taylor (1950) formulated it, seems to belong to a fairly immediate, general criterion of organizational effectiveness (about Cell 15).

### The Problem of Contamination-Deficiency

The problem of contamination and deficiency is most urgent at the level of organizational effectiveness. We can include unwanted sources of variance in our organizational measures, as, for example, ethnic, racial, political, and familial preferences, and we can exclude other, particularly long-term sources, such as planning for the future R&D efforts. (One firm actually forbids the purchase of capital equipment unless it can be paid for out of annual profits in that operating unit!) To the extent to which contamination occurs, the criterion measure tends to drift off the chart shown in Figure 1. Deficiency leads it toward the immediate and specific (try Cell 16). This problem is what Astin (1964) called the problem of conceptual criteria.

In one sense, this problem is the central problem of criteria. If operational decisions are based on contaminated or deficient grounds, no degree of care in the results or behaviors segments will preserve organizational effectiveness. The operational problem here is that we have nothing to lean on but administrative judgment, which can build an Edsel out of pieces of reasonably sound criterion research.

### Validation by the Multitrait-Multimethod Matrix

The validity of a particular method of measuring a criterion variable may be determined relatively efficiently by evaluation of convergent and discriminant validity (Campbell & Fiske, 1959). Briefly, a measure should agree with other measures of the same trait more closely than with measures using the same method of measurement designed to measure different traits. In factor analytic terms, the measure should load more highly on a trait factor than on a method factor. The factor approach in my opinion is less susceptible to the effects of small fluctuations in the size of correlation coefficients than the Campbell-Fiske model. Either the inspection of the correlation matrix or factor analysis will evaluate the extent to which a measure is central to the concept being evaluated. Either requires the time to take multiple measures using a

variety of methods, which will not be popular with management.

Most criteria have not been validated at all; they have been established by fiat. Convergence and discrimination seem to be minimal requirements before an entire decision-making process is to be constructed on the basis of a measure. A little salesmanship by the psychologist to management is in order here.

### The Dynamic Nature of Criteria

The time dimension in Figure 1 is an important one. The "same" measure has different factorial structure at different points in time (Fleishman & Hempel, 1954; Fleishman & Fruchter, 1960). Ghiselli (1956) has emphasized that relations between tests and criteria are not necessarily stable, and that changes in performance occur over very long periods of time.

Longitudinal studies of criteria in actual practice are in order. This is one task for psychologists in industry who can undertake longitudinal projects, and who have the courage to sell them to management. Only when we have real-life studies of the relative stability of relationships between treatments or predictors and criteria over time can we know when to generalize from a single study to a general problem. We live in a time in which it is necessary to use law to enforce even one-shot test validations, and in which experimenters use college sophomores in one-hour laboratory experiments to generalize to the world of the long working life of people. Studies such as those of Fleishman and Fruchter (1960) on telegraphers in which the structure of abilities changed with both treatments and abilities need to be extended. There, the actual observed behaviors shifted with time and, consequently, so did the relationship of a (various) predictor(s) with a criterion (criteria).

Managers, with the help of psychologists, need to decide at what stage(s) in progress measures should be taken as criteria. This decision should invoke all of the considerations involved in the time dimension.

### Individual Styles

The results slab involves the combination of behaviors into results. This combination is complicated by the fact that different people achieve an end of the same value by a different combination of behaviors. Ghiselli (1956) emphasized this problem of personal styles in limiting prediction of performance by a single regression equation of tests (or behaviors) against a single criterion measure. One manager achieves production by a strong emphasis on human relations, for example, while another emphasizes production control. Ghiselli implies that subjects be separated according to styles and that prediction be attempted only within a given style. This procedure, unfortunately, greatly diminishes the *N*. It holds the promise, however, of raising our obtained relationships.

### Use of Criteria in Counseling

Development of employees (or managers) by feeding back the strong and weak points in their performance as indicated by criterion measures and devising plans for improvement has been one of the explicit aims of top management. Yet, the procedure is seldom used. One of the reasons is that criteria developed for administrative purposes are personally very threatening and discussion of them leads to defensiveness and deflation (Meyer, Kay, & French, 1965). Defensiveness can be reduced by making sure the person evaluated can participate in planning for his own improvement (Meyer et al., 1965; French, Kay, & Meyer, 1966; Bassett & Meyer, 1968). But, it seems desirable to separate the salary evaluation aspects of rating from the counseling aspects.

Successful combination of administrative (promotional) and counseling interviews has been reported in an assessment center (Acker & Perlson, 1971) perhaps because these applicants were volunteers for the program. At any rate, counseling, to an even greater degree than research, requires that multiple criteria be gathered which are psychologically discriminable and capable of being communicated effectively.



### How to Determine an Organizational Criterion

The procedure for developing a criterion has been clearly summarized by Guion:

1. Analyze the job and/or the organizational needs by new, yet-to-be-developed techniques.
2. Develop measures of actual behavior relative to the behavior expected, as identified in job and need analysis. These measures are to supplement measures of the consequences of work—the so-called objective criteria commonly tried at present.
3. Identify the criterion dimensions underlying such measures by factor analysis or cluster analysis or pattern analysis.
4. Develop reliable measures, each with high construct validity, of the elements so identified.
5. For each independent variable (predictor), determine its predictive validity for *each one* of the foregoing criterion measures, taking them one at a time. (Guion, 1961, p. 148)

This procedure allows judgment concerning weighting of different criteria *after* the empirical data are in—finally allowing some weighting according to predictability. Movement to the organizational "slab" occurs only when behaviors and results have been analyzed.

## RECURRENT PROBLEMS

### Relevance and Reliability of Observation and Rating

Observation and interpretation hold the key to establishment of effective criteria. These are related to the convergence and discriminability of both the observational measures and the combined rating that results from those observations.

Establishing the reliability of judgments is essential, as is the correlation of behaviors with results and results with organizational effectiveness. These steps have seldom been performed; investigators would improve their effectiveness if they would do so.

### Who Should Evaluate?

As our discussion has indicated, most criteria boil down to ratings. The crucial recurrent problem is who should rate. The first level requires that those people rate who have observed actual behavior, which strongly suggests peer ratings. The second level requires inferences from the first, which suggests supervisors' ratings. The third level is managerial, and is poorly represented in our diagram. Here the psychologist can help pull the results together, but the final decision about how to use such judgments belongs to top management.

### Multidimensionality and the Use of Multiple Measures

The search for *the* elusive criterion continues despite strong evidence that it does not exist. The fact of multidimensionality poses a practical problem at the level of organizational effectiveness. Various objectives have to be weighted in making decisions. Organizational goals have to be spelled out, including community responsibilities. Nowhere is this problem more evident than in the area of minority employment (see Chapter 18), where community goals are often in conflict with economic goals. Executive decisions have to be made.

Criteria for different purposes need to be separated. They should parallel the predictors or treatments in generality and immediacy. More than one measure should be taken if possible at the required level so that convergence can be evaluated.

### Equivalence of Measures

Finally, one measure cannot be freely substituted for another without establishing equivalence, which means much more than

high intercorrelation. It involves similar correlations with other variables and similar responses to treatments (see Gulliksen, 1950; Smith, Kendall, & Hulin, 1969, pp. 152–158; Smith, Smith, Baumgartel, Gliner, & Goodale, 1971). Thoughtless substitution of one rating for another or one measure of absences for another can greatly disrupt relationships with other variables and hence resulting decisions.

## TOWARD AN INTEGRATION OF CRITERION MODELS

Three criterion models and the possibility of their integration through construct validation have been discussed by James (1973). Since the models discussed by James are the same as the ones presented in this chapter, a brief review of his comments and suggestions serves a useful summarizing role for what has been said here.

His discussion focuses on three criterion models: (1) the ultimate criterion model [as developed by Thorndike (1949) and as presented in this chapter]; (2) the multiple criterion model [as articulated by Ghiselli (1956), Guion (1961), Dunnette (1963), Wallace (1965), and Schmidt and Kaplan (1971) and discussed in this chapter]; and (3) the general criterion model represented by the view of managerial effectiveness presented by Campbell, Dunnette, Lawler, and Weick (1970). The last model differs from the multiple model only to the extent of specifying the major components and probable internal dynamics of the multiple measures required in efforts to understand performance effectiveness. In this sense, I have implied use of the general model throughout this chapter by specifying the various types and levels of measurement that may be desirable in any concerted effort to measure the facets of job performance.

The major thrust of James's argument is that the three models can best be melded into an integrated approach to analysis of criteria through the theory and technology of construct validation (Cronbach & Meehl, 1955). This follows directly the counsel of Kavanagh, MacKinney, and Wolins (1971) when they state that since the ultimate criterion "can best be described as a psychological construct . . . the process of determining the relevance of the immediate to the ultimate criterion becomes one of construct validation" (p. 35). The case for construct validation as a central point of methodological and theoretical emphasis in criterion research and development is easy to make. Construct validation is the method of choice because it is the only way of understanding *what* is being measured by criteria which cannot be validated by the traditional methods of empirical validation. In particular, as pointed out by James (1973), "the need to identify criterion constructs becomes crucial whenever contaminated measures, such as ratings, are employed, especially multiple ratings from different sources, or whenever operationally defined objective criteria (typically global) are not available" (p. 79). In other words, an orientation toward construct validation in criterion research is the best way of guarding against a hopelessly incomplete job of criterion development. In essence, *all* the possible sources and measures of performance variation discussed in this chapter need to be given an opportunity to be studied within the nomological net of each ultimate criterion construct. These should include measures sampling all degrees of complexity, all relevant time periods, and all levels of measurement—behavioral, results, and organizational consequences. Coverage should include but not be restricted to such measures as behaviorally based performance ratings developed by and completed by several raters; objective measures of job performance, ability, motivation-satisfaction morale; measures of situational parameters; and global measures of organizational outcomes. A full and complete understanding of the ultimate constructs of performance effectiveness in any specific job-person-organization setting can best be gained by an ongoing and continuing program of construct validation.



## REFERENCES


Acker, S. R., & Perlson, M. R. Can we sharpen our management of human resources? *Behavioral Sciences Applications*, Corporate Personnel Department, Olin Corporation, 1971.

Argyle, M., Gardner, G., & Cioffi, F. Supervisory methods related to productivity, absenteeism, and labour turnover. *Human Relations*, 1958, 11, 23–40.

Astin, A. W. Criterion-centered research. *Educational and Psychological Measurement*, 1964, 24, 807–822.

Barnes, R. M. *Motion and time study*. (5th ed.) New York: Wiley, 1963.

Bass, B. M. Reducing leniency in merit ratings. *Personnel Psychology*, 1956, 9, 359–369.

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

Bassett, G. A., & Meyer, H. H. Performance appraisal based on self-review. *Personnel Psychology*, 1968, 21, 421–430.

Baumgartel, H., & Sobol, R. Background and organizational factors in absenteeism. *Personnel Psychology*, 1959, 12, 431–443.

Bayroff, A. G., Haggerty, H. R., & Rundquist, E. A. Validity of ratings as related to rating techniques and conditions. *Personnel Psychology*, 1954, 7, 93–114.

Behrend, H. Absence and turnover in a changing economic climate. *Occupational Psychology*, 1953, 27, 69–79.

Bentz, V. J. The Sears experience in the investigation, description, and prediction of executive behavior. In *Predicting managerial success*. Ann Arbor, Mich.: Foundation for Research in Human Behavior, 1968, 59–152.

Berkshire, J. R., & Highland, R. W. Forced-choice performance rating. *Personnel Psychology*, 1953, 6, 355–378.

Besco, R. O., & Lawshe, C. H. Foreman leadership as perceived by superiors and subordinates. *Personnel Psychology*, 1959, 12, 573–582.

Besnard, G. G., & Briggs, L. J. Measuring job proficiency by means of a performance test. In E. A. Fleishman (Ed.), *Studies in personnel and industrial psychology*. (Rev. ed.) Homewood, Ill.: Dorsey Press, 1967.

Biesheuvel, S. Personnel selection. *Annual Review of Psychology*, 1965, 16, 295–324.

Bingham, W. V., & Davis, W. T. Intelligence test scores and business success. *Journal of Applied Psychology*, 1924, 8, 1–22.

Blankenship, A. B., & Taylor, H. R. Prediction of vocational proficiency in three machine operations. *Journal of Applied Psychology*, 1938, 22, 518–526.

Blum, M. L., & Naylor, J. C. *Industrial psychology: Its theoretical and social foundations*. New York: Harper and Row, 1968, Chapter 7.

Blumberg, H. H., DeSoto, C. B., & Kuethe, J. L. Evaluation of rating scale formats. *Personnel Psychology*, 1966, 19, 243–259.

Borman, W. C. The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance*, 1974, 12, 105–124.

Borman, W. C., & Dunnette, M. D. *Selection of components to comprise a Naval Personnel Status Index (NPSI) and a strategy for investigating their relative importance*. Minneapolis: Personnel Decisions, 1974.

Bray, D. W., & Campbell, R. J. Selection of salesmen by means of an assessment center. *Journal of Applied Psychology*, 1968, 52, 36–41.

Bray, D. W., & Moses, J. L. Personnel selection. *Annual Review of Psychology*, 1972, 23, 545–576.

Brogden, H. E., & Taylor, E. K. The dollar criterion: Applying the cost accounting concept to criterion construction. *Personnel Psychology*, 1950, 3, 133–154.

Buel, W. D. The validity of behavioral scale items for the assessment of individual creativity. *Journal of Applied Psychology*, 1960, 44, 407–412.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81–105.

Campbell, J. P., Dunnette, M. D., Lawler, E. E. III, & Weick, K. E. Jr. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Champney, H. The measurement of parent behavior. *Child Development*, 1941, 12, 131–166.

Christal, R. E. JAN: A technique for analyzing group judgment. *The Journal of Experimental Education*, 1968, 36, 24–27.

Cliff, N. Adverbs as multipliers. *Psychological Review*, 1959, 66, 27–44.
Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 62, 281–302.
Daniels, H. W., & Edgerton, H. A. The development of criteria of safe operation for groups. *Journal of Applied Psychology*, 1954, 38, 47–53.
Doll, R. E., & Longo, A. A. Improving the predictive effectiveness of peer ratings. *Personnel Psychology*, 1962, 15, 215–220.
Dunnette, M. D. A note on the criterion. *Journal of Applied Psychology*, 1963, 47, 251–254.
Dunnette, M. D. *Personnel selection and placement*. Belmont, Calif.: Wadsworth Publishing, 1966.
Dunnette, M. D., Campbell, J. P., & Hellervik, L. W. *Job behavior scales for Penney Co. department managers*. Minneapolis: Personnel Decisions, 1968. (Cited in J. P. Campbell, M. D. Dunnette, E. E. Lawler III, & K. E. Weick Jr., *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970, 119–123.)
Edgerton, H. A., & Kolbe, L. E. The method of minimum variation for the coordination of criteria. *Psychometrika*, 1936, 1, 185–187.
Edwards, A. E. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
Ewart, E. S., Seashore, S. E., & Tiffin, J. A factor analysis of an industrial merit rating scale. *Journal of Applied Psychology*, 1941, 25, 481–486.
Ewen, R. B. Weighting components of job satisfaction. *Journal of Applied Psychology*, 1967, 51, 63–73.
Ferguson, L. W. The development of a method of appraisal for assistant managers. *Journal of Applied Psychology*, 1947, 31, 306–311.
Finkle, R. B., & Jones, W. S. *Assessing corporate talent*. New York: Wiley, 1970.
Fishbein, M., Ed. *Readings in attitude theory and measurement*. New York: Wiley, 1967.
Fiske, D. W. Values, theory, and the criterion problem. *Personnel Psychology*, 1951, 4, 93–98.
Flanagan, J. C. Critical requirements: A new approach to employee evaluation. *Personnel Psychology*, 1949, 2, 419–425.
Flanagan, J. C. The critical incident technique. *Psychological Bulletin*, 1954, 51, 327–355.
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., & Hempel, W. E. Jr. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 1954, 18, 239–252.
Fleishman, E. A., Harris, E. F., & Burtt, H. E. Leadership and supervision in industry. *Bureau of Education Monographs* # 33. Columbus: Ohio State University, 1955.
Fogli, L., Hulin, C. L., & Blood, M. R. Development of first-level behavioral criteria. *Journal of Applied Psychology*, 1971, 55, 3–8.
Forehand, G. A. Assessments of innovative behavior: Partial criteria for the assessment of executive performance. *Journal of Applied Psychology*, 1963, 47, 206–213.
Frederiksen, N., Saunders, D. R., & Ward, B. The in-basket test. *Psychological Monographs*, 1957, 71, No. 9 (Whole No. 438).
French, J. R. P. Jr., Kay, E., & Meyer, H. H. Participation and the appraisal system. *Human Relations*, 1965, 18, 3–20.
French, J. W. The validity of some objective personality tests for a leadership criterion. *Educational and Psychological Measurement*, 1954, 14, 34–49.
Funk & Wagnalls. *Standard college dictionary*. (Text ed.) New York: Harcourt, Brace, & World, 1963.
Ghiselli, E. E. Dimensional problems of criteria. *Journal of Applied Psychology*, 1956, 40, 1–4.
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.
Ghiselli, E. E., & Brown, C. W. *Personnel and industrial psychology*. New York: McGraw-Hill, 1955.
Ghiselli, E. E., & Haire, M. The validation of selection tests in the light of the dynamic character of criteria. *Personnel Psychology*, 1960, 13, 225–231.
Gifford, W. W. Does business want scholars? *Harper's Magazine*, 1928, 156, 669–674.
Gordon, L. V., & Medlund, F. F. The cross-group stability of peer ratings of leadership potential. *Personnel Psychology*, 1965, 18, 173–177.
Grant, D. L. A factor analysis of managers' ratings. *Journal of Applied Psychology*, 1955, 39, 283–286.
Grant, D. L., & Bray, D. W. The assessment center in the measurement of potential for business management. *Psychological Monographs*, 1966, 80 (Whole No. 625).
Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954, 274–278.
Guion, R. M. Criterion measurement and personnel judgments. *Personnel Psychology*, 1961, 14, 141–149.
Guion, R. M. *Personnel testing*. New York: McGraw-Hill, 1965.
Guion, R. M. Personnel selection. *Annual Review of Psychology*, 1967, 18, 191–216.
Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.
Hakel, M. D. Jr. Perceiver differences in interpersonal perceptions: An analysis of interrater agreement on scaled-expectation rating scales in an employment interview setting. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, 1966.
Hartshorne, H., & May, M. A. *Studies in service and self-control*. New York: Macmillan, 1929.
Heise, D. R. Some methodological issues in semantic differential research. *Psychological Bulletin*, 1969, 72, 406–422.
Henry, W. E. Executive personality and job success. *AMA Personnel Series*, 1948, #120.
Hollander, E. P. Buddy ratings: Military research and industrial implications. *Personnel Psychology*, 1954, 7, 385–395.
Hollander, E. P. The friendship factor in peer nominations. *Personnel Psychology*, 1956, 9, 425–447.
Hollander, E. P. The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology*, 1957, 41, 85–90.
Hollander, E. P. Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology*, 1965, 49, 434–438.
Howe, E. S. Further comparisons of two short-form derivations of the Taylor manifest anxiety scale. *Psychological Reports*, 1960, 6, 21–22.
Howe, E. S., & Silverstein, A. B. Comparison of two short-form derivatives of the Taylor manifest anxiety scale. *Psychological Reports*, 1960, 6, 9–10.
Hulin, C. L. The measurement of executive success. *Journal of Applied Psychology*, 1962, 46, 303–306.
Huse, E. F., & Taylor, E. K. Reliability of absence measures. *Journal of Applied Psychology*, 1962, 46, 159–160.
Izard, B. R., & Rosenberg, S. Effectiveness of a forced-choice leadership test under varied experimental conditions. *Educational and Psychological Measurement*, 1958, 18, 57–62.
James, L. R. Criterion models and construct validity for criteria. *Psychological Bulletin*, 1973, 80, 75–83.
Jaques, E. *Equitable payment*. New York: Wiley, 1961.
Jurgensen, C. E. A fallacy in the use of median scale values in employee check lists. *Journal of Applied Psychology*, 1949, 33, 56–58.
Katzell, R. A. Industrial psychology. *Annual Review of Psychology*, 1957, 8, 237–268.
Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, 1971, 75, 34–49.
Kay, B. R. The use of critical incidents in a forced-choice scale. *Journal of Applied Psychology*, 1959, 43, 269–270.
Kendall, L. M., & Hilton, T. L. Rationale and results of an attempt to develop behaviorally anchored rating criteria for students in graduate schools of business administration. Paper presented at 73rd American Psychological Association Convention, September 5, 1965.
Kendall, W. E. Industrial psychology. *Annual Review of Psychology*, 1955, 6, 217–250.
Kerr, W. A., Koppelmeir, G., & Sullivan, J. J. Absenteeism, turnover, and morale in a metals fabrication factory. *Occupational Psychology*, 1951, 25, 50–55.
Kirchner, W. K. Relationships between supervisory and subordinate ratings of technical personnel. *Journal of Industrial Psychology*, 1966, 3, 57–60.
Kirchner, W. K., & Dunnette, M. D. Using critical incidents to measure job proficiency factors. *Personnel*, 1957, 34, 54–59.
Kirchner, W. K., & Reisberg, D. J. Differences between better and less effective supervisors in appraisal of subordinates. *Personnel Psychology*, 1962, 15, 295–302.
Klores, M. S. Rater bias in forced-distribution performance ratings. *Personnel Psychology*, 1966, 19, 411–421.
Knauft, E. B. Construction and use of weighted check list rating scales for two industrial

situations. *Journal of Applied Psychology*, 1948, 32, 63–70.
Koltuv, B. B. Some characteristics of intra-judge trait intercorrelations. *Psychological Monographs*, 1962, 76, 33 (Whole No. 552).
Kornhauser, A. W. A statistical study of a group of specialized office workers. *Journal of Personnel Research*, 1923, 2, 103–123.
Krick, E. V. *Methods engineering*. New York: Wiley, 1962.
Kunin, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, 8, 65–78.
Kurtz, A. B. The simultaneous prediction of any number of criteria by the use of a unique set of weights. *Psychometrika*, 1937, 2, 95–101.
Lawler, E. E. III. The multi-trait-multi-rater approach to measuring managerial job performance. *Journal of Applied Psychology*, 1967, 51, 369–381.
Lawshe, C. H., Kephart, N. C., & McCormick, E. J. The paired comparison technique for rating performance of industrial employees. *Journal of Applied Psychology*, 1949, 33, 69–77.
Lefkowitz, J. Effect of training on the productivity and tenure of sewing machine operators. *Journal of Applied Psychology*, 1970, 54, 81–86.
Lifson, K. A. Errors in time-study judgments of industrial work pace. *Psychological Monographs*, 1953, 67, No. 5 (Whole No. 358).
Likert, R., & Seashore, S. E. Making cost control work. *Harvard Business Review*, 1963, 41, 96–108.
Lyons, T. F. Turnover and absenteeism: A review of relationships and correlates. *Personnel Psychology*, 1972, 25, 271–281.
Maas, J. B. Patterned scale expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology*, 1965, 49, 431–433.
McGehee, W. Cutting training waste. *Personnel Psychology*, 1948, 1, 331–340.
McGehee, W., & Owen, E. B. Authorized and unauthorized rest pauses in clerical work. *Journal of Applied Psychology*, 1940, 24, 605–614.
MacKinney, A. C. An assessment of performance change: An inductive example. *Organizational Behavior and Human Performance*, 1967, 2, 56–72.
Madden, J. M. Comparison of three methods of rating-scale construction. *Journal of Industrial Psychology*, 1964, 2, 43–50.
Madden, J. M., & Bourdon, R. D. Effects of variations in rating scale format on judgment. *Journal of Applied Psychology*, 1964, 48, 147–151.
Maguire, T. O., & Glass, G. V. Component profile analysis (COPAN)—An alternative to PROF. *Educational and Psychological Measurement*, 1968, 28, 1021–1033.
Mann, F. C., & Dent, J. K. The supervisor: Member of two organizational families. *Harvard Business Review*, 1954, 32, 103–112.
Marks, M. R. Review of Ronan, W. W., & Prien, E. P., Toward a criterion theory: A review and analysis of research and opinion. *Personnel Psychology*, 1967, 20, 216–218.
Metzner, H., & Mann, F. Employee attitudes and absences. *Personnel Psychology*, 1953, 6, 467–485.
Meyer, H. H. The validity of the in-basket test as a measure of managerial performance. *Personnel Psychology*, 1970, 23, 297–307.
Meyer, H. H., Kay, E., & French, J. R. P. Jr. Split roles in performance appraisal. *Harvard Business Review*, 1964, 43, 124–129.
Mordkoff, A. M. Functional versus nominal autonomy in semantic differential scales. *Psychological Reports*, 1965, 16, 691–692.
Mueser, R. E. The weather and other factors influencing employee punctuality. *Journal of Applied Psychology*, 1953, 37, 329–337.
Naylor, J. C., & Wherry, R. J. Sr. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement*, 1965, 25, 969–986.
Obradović, J. Modification of the forced-choice method as a criterion of job proficiency. *Journal of Applied Psychology*, 1970, 54, 228–233.
Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.
Otis, J. L. The criterion. In W. H. Stead, C. Shartle, & Associates, *Occupational counseling techniques*. New York: American Book Co., 1940.
Owens, W. A., & Jewell, D. O. Personnel selection. *Annual Review of Psychology*, 1969, 20, 419–446.
Passini, F. T., & Norman, W. T. A universal conception of personality structure? *Journal of Personality and Social Psychology*, 1966, 4, 44–49.
Passini, F. T., & Norman, W. T. Ratee relevance in peer nominations. *Journal of Applied Psychology*, 1969, 53, 185–187.
Porter, L. W., & Steers, R. M. Organizational, work, and personal factors in employee turnover and absenteeism. *Journal of Applied Psychology*, 1973, 80, 151–176.
Prien, E. P. Dynamic character of criteria: Organizational change. *Journal of Applied Psychology*, 1966, 50, 501–504.
Prien, E. P., & Kult, M. Analysis of performance criteria and comparison of a priori and empirically derived keys for a forced-choice scoring. *Personnel Psychology*, 1968, 21, 505–513.
Rambo, W. W. The construction and analysis of a leadership behavior check list for industrial managers. *Journal of Applied Psychology*, 1958, 42, 409–415.
Richardson, M. W., & Kuder, G. V. Making a rating scale that measures. *Personnel Journal*, 1933, 12, 36–40.
Roach, D. E., & Wherry, R. J. Performance dimensions of multi-line insurance agents. *Personnel Psychology*, 1970, 23, 239–250.
Roadman, H. E. An industrial use of peer ratings. *Journal of Applied Psychology*, 1964, 48, 211–214.
Roethlisberger, F. J., & Dickson, W. J. *Management and the worker*. Cambridge: Harvard University Press, 1938.
Ronan, W. W. A factor analysis of eight job performance measures. *Journal of Industrial Psychology*, 1963, 1, 107–112.
Ronan, W. W., & Prien, E. P. *Towards a criterion theory: A review and analysis of research and opinion*. Greensboro, N.C.: The Richardson Foundation, 1966.
Ross, P. F. Reference groups in man-to-man job performance rating. *Personnel Psychology*, 1966, 19, 115–142.
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 a hypothesis regarding the "restriction of output." *Journal of Applied Psychology*, 1946, 30, 320–327. (b)
Rothe, H. F. Output rates among machine operators: I. Distributions and their reliability. *Journal of Applied Psychology*, 1947, 31, 384–389.
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.
Rothe, H. F., & Nye, C. T. Output rates among machine operators: III. A nonincentive situation in two levels of business activity. *Journal of Applied Psychology*, 1961, 45, 50–54.
Rotter, G. S., & Tinkleman, V. Anchor effects in the development of behavior rating scales. *Educational and Psychological Measurement*, 1970, 30, 311–318.
Rush, C. H. Jr. A factorial study of sales criteria. *Personnel Psychology*, 1953, 6, 9–24.
Ryan, T. A. *Work and effort*. New York: Ronald Press, 1947.
Ryan, T. A., & Smith, P. C. *Principles of industrial psychology*. New York: Ronald Press, 1954, 46–58.
Salvendy, G., Seymour, W. D., & Corlett, E. N. Comparative study of static versus dynamic scoring of performance tests for industrial operators. *Journal of Applied Psychology*, 1970, 54, 135–139.
Schmidt, F. R., & Kaplan, L. B. Composite versus multiple criteria: A review and resolution of the controversy. *Personnel Psychology*, 1971, 24, 419–434.
Schneider, D. E., & Bayroff, A. G. The relationship between rater characteristics and validity of ratings. *Journal of Applied Psychology*, 1953, 37, 278–280.
Schuh, A. J. The predictability of employee tenure: A review of the literature. *Personnel Psychology*, 1967, 20, 133–152.
Schultz, D. G., & Siegel, A. I. Generalized Thurstone and Guttman scales for measuring technical skills in job performance. *Journal of Applied Psychology*, 1961, 45, 137–142.
Schultz, D. G., & Siegel, A. I. The analysis of job performance by multidimensional scaling techniques. *Journal of Applied Psychology*, 1964, 48, 329–335.
Scott, W. A. Comparative validities of forced-

choice and single-stimulus tests. *Psychological Bulletin*, 1968, 70, 231–244.

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

Seashore, S. E., & Yuchtman, E. Factorial analysis of organizational performance. *Administrative Science Quarterly*, 1967, 12, 377–395.

Sharon, A. T., & Bartlett, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology*, 1969, 22, 251–263.

Siegel, A. I., & Pfeiffer, M. G. Factorial congruence in criterion development. *Personnel Psychology*, 1965, 18, 267–280.

Sisson, E. D. Forced-choice: The new Army rating. *Personnel Psychology*, 1948, 1, 365–381.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, 6, 649–744.

Smith, O. W., & Landy, F. Grid versus graphic scaling of importance and presence of some college experiences. *Perceptual and Motor Skills*, 1969, 29, 146.

Smith, O. W., Smith, P. C., Baumgartel, R., Gliner, J., & Goodale, J. Psychology of the scientist: XXX. Replication: What is it? *Perceptual and Motor Skills*, 1971, 33, 691–697.

Smith, P. C. The curve of output as a criterion of boredom. *Journal of Applied Psychology*, 1953, 37, 69–74.

Smith, P. C., & Gold, R. A. Prediction of success from examination of performance during the training period. *Journal of Applied Psychology*, 1956, 40, 83–86.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, 47, 149–155.

Smith, P. C., & Lem, C. Positive aspects of motivation in repetitive work: Effects of lot size upon spacing of voluntary work stoppages. *Journal of Applied Psychology*, 1955, 39, 330–333.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.

Smith, P. C., & Taylor, J. G. An investigation of the shape of learning curves for industrial motor tasks. *Journal of Applied Psychology*, 1956, 40, 142–149.

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

Stockford, L., & Bissell, H. W. Establishing a graphic-rating scale. In W. E. Fleishman (Ed.), *Studies in personnel and industrial psychology*. (Rev. ed.) Homewood, Ill.: Dorsey Press, 1967.

Taylor, E. K., & Nevis, E. C. Personnel selection. *Annual Review of Psychology*, 1961, 12, 403–405.

Taylor, E. K., & Wherry, R. J. A study of leniency in two rating systems. *Personnel Psychology*, 1951, 4, 39–47.

Taylor, E. K., Schneider, D. E., & Clay, H. Short forced-choice ratings work. *Personnel Psychology*, 1954, 7, 245–252.

Taylor, E. K., Schneider, D. E., & Symons, N. A. A short forced-choice evaluation form for salesmen. *Personnel Psychology*, 1953, 6, 393–401.

Thomson, H. A. Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. *Journal of Applied Psychology*, 1970, 54, 496–502.

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

Thornton, G. C. The relationship between supervisory- and self-appraisals of executive performance. *Personnel Psychology*, 1968, 21, 441–455.

Tiffin, J., & Phelan, R. F. Use of the Kuder preference record to predict turnover in an industrial plant. *Personnel Psychology*, 1953, 6, 195–204.

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

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Travers, R. M. W. A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin*, 1951, 48, 62–70.

Tucker, M. F., Cline, V. B., & Schmitt, J. R. Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. *Journal of Applied Psychology*, 1967, 51, 131–138.

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

Uhrbrock, R. S. Standardization of 724 rating scale statements. *Personnel Psychology*, 1950, 3, 285–316.

Uhrbrock, R. S. 2,000 scaled items. *Personnel Psychology*, 1961, 14, 375–420.

Vielhaber, D. P., & Gottheil, E. First impressions and subsequent ratings of performance. *Psychological Reports*, 1965, 17, 916.

Viteles, M. S. General review and summary: Five years of research, a summary of outcomes. *Psychological Bulletin*, 1945, 42, 489–526.

Wallace, S. R. Criteria for what? *American Psychologist*, 1965, 20, 411–417.

Wallace, S. R., & Weitz, J. Industrial psychology. *Annual Review of Psychology*, 1955, 6, 217–250.

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

Weitz, J., & Nuckols, R. C. A validation study of "How supervise?" *Journal of Applied Psychology*, 1953, 37, 7–8.

Wherry, R. J., & Fryer, D. H. Buddy ratings: Popularity contest or leadership criteria? *Personnel Psychology*, 1949, 2, 147–159.

Wherry, R. J. Sr., & Naylor, J. C. Comparison of two approaches—JAN and PROF—for capturing rater strategies. *Educational and Psychological Measurement*, 1966, 26, 267–286.

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.

Wiley, L. Relation of characteristics ratings to performance ratings. *Journal of Industrial Psychology*, 1964, 2, 7–15.

Wiley, L., & Jenkins, W. S. Selecting competent raters. *Journal of Applied Psychology*, 1964, 48, 215–217.

Williams, J. D., Harlow, S. D., Lindem, A., & Gab, D. A judgment analysis program for clustering similar judgmental systems. *Educational and Psychological Measurement*, 1970, 30, 171–173.

Yuzuk, R. P. The assessment of employee morale. Columbus: The Ohio State University Bureau of Business Research, 1961, Monograph No. 99.

Zavala, A. Development of the forced-choice rating scale technique. *Psychological Bulletin*, 1965, 63, 117–124.