1986

10. Needed Directions for Measurement in Work Settings

Mary L. Tenopyr
The American Telephone and Telegraph Company

10 Needed Directions for Measurement in Work Settings

Mary L. Tenopyr
The American Telephone and Telegraph Company

One of the greatest needs in psychology today is the establishment of more rigorous psychological measurement practices in the millions of work settings throughout the country. Today, any semblance of precise measurement appears to be limited to the largest of employers. Only the biggest corporations and the major governmental units, such as those in the federal government, have the scientific staffs to conduct the research and development work necessary to provide the type of measurement that is so needed.

Psychological measurement in work settings has a profound effect upon American society. Indeed, it affects almost all citizens' lives. Employees, job seekers, and their families all, to some extent, have their lives shaped by the psychological measurement practices of employers. How a breadwinner is appraised in a job application or a performance evaluation situation may have an impact on many lives. What job one works in, and even whether one works at all, are all decided mainly on the basis of some psychological measurement, however imperfect. The indirect effects of measurement also must be considered; many of those who have power over us, e.g., supervisors or government officials, were measured in some way when they were selected for their jobs, and also, they remain in their jobs as a result of some application of measurement.

The implications of psychological measurement in the workplace for the educational system cannot be lightly dismissed. Obviously, one major function of education is to prepare students for work and careers. Only through measurement in employment settings can the critical abilities and skills necessary to develop educational curricula be designated and defined. Only then can students be adequately prepared.

The relationship between psychological measurement and the economic health of the country is more nebulous, but probably should be considered to be

more than nominal. The well-documented productivity declines of the 1970s were not entirely explainable by typical economic measures, such as amount invested in research and development (Dennis, 1979). The productivity of the individual worker may well have been partially responsible for this decline, and hence, by implication, the methods by which he or she was selected for and retained in the job may well have played a part.

One may well ask why, if measurement in the workplace has so many potential effects in our society, has it not been a subject of great concern in employing organizations. The answers do not emerge readily. There is probably no single explanation for the general lack of precise measurement in the employing community. Certainly, the legal climate for measurement is considered inhospitable by many employers. Results of a recent survey (Bureau of National Affairs, 1983) indicate that the little employee selection testing which has been going on is on the decline. It appears that about 5% to 9% of employers are doing any testing at all. Employers who are dropping testing have indicated that they are doing so because of fear of litigation. However, fear of legal difficulties is only a part of the story. The abuses of testing in business several decades ago became part of American folklore, mainly as a result of the activities of popular writers (Hoffman, 1962; Whyte, 1956). Despite the fact that the lay criticism was mainly of personality inventories, a dark cloud fell over all testing by employers. Many business people began to speak of testing in terms usually reserved for activities such as examining the entrails of birds. Unfortunately, those employers who did continue testing often did so without benefit of validation research. This type of testing culminated in a U.S. Supreme Court decision (*Griggs v. Duke Power Company,* 1971), which mandated a demonstration of job relatedness for any test having a disparate impact upon a minority group. The response to this decision and the many court decisions and administrative actions that have followed was two-fold. Most employers, troubled by the bad reputation of testing, coupled with the possibility of legal difficulties, fled from testing. At the other extreme, a few major employers began utilizing testing research staffs and tried to meet the provisions of the law. Thus, the situation we have today with less than 10% of employers testing (Bureau of National Affairs, 1983) has come to prevail. Most employees are selected by interviews and reference checks, both of which are usually of uncertain validity.

MAJOR CONSIDERATIONS

One of the most fruitful new directions that can be taken, relative to measurement in work settings, is to undertake a massive educational program, not only for those responsible for employment procedures, but also for those who make government policy and law. However, we must concurrently take some actions to ensure that our scientific house is in order. In fact, what is needed is a synergistic combination of educational and scientific considerations. For exam-

ple, the research and development funds necessary for the scientific achievements we need are most logically supplied by employers, but this money will not be furnished unless employers recognize the value of sound psychological measurement. In particular, the relative merits of various alternate forms of measurement must become common knowledge in the employing and governmental communities. Reilly and Chao (1982) have pointed out that no alternatives to traditional tests are more valid, and most of them are less valid. A related important need is to help frame government policies so that the standards for use of tests are not so rigorous, that even more tests are abandoned in favor of techniques like unvalidated, unstructured interviews.

Also, those responsible for funding need to be aware that the development of reliable and valid measuring methods is not inexpensive. Concomitantly, these policy makers must become aware of the potential utility of sound measurement for increasing performance and productivity. In other words, these persons must come to know that the return on investment in sound measurement is usually substantial.

A third educational objective is to teach employers to recognize the difference between responsible experts in measurement and those with lesser skill or those who recommend measurement programs not based on sound research. It is the author's opinion that many of the difficulties employment testing faces today could have been averted if, in the past, employers had been trained to evaluate recommendations for testing programs on their merits instead of being unduly influenced by the salesmanship of those who proposed such programs.

Coupled with education, there are a number of scientific considerations that deserve attention. Although science should never be frozen in time, one cannot conduct an educational program relative to a scientific endeavor unless there are coherent principles underlying the science. There are a number of needs for research and development that would make the principle base for measurement more supportive. First, there needs to be a conceptualization of validity which is applicable in employment settings. Second, appropriate systems of constructs are required. Paralleling this need, is a need to reduce work requirements into meaningful and manageable dimensions; we need taxonomies of both abilities and work. Third, is a need for clarification regarding job analysis which is one of the major ancillaries to measurement. A fourth need is for performance measurement techniques which are reflective of performance and, at the same time, feasible to apply. Fifth, is a need for guidance in the development of alternatives to traditional paper-pencil tests, such as interviews and work samples. Finally, there is a need for clarification of the differential prediction area. In particular, there needs to be a meshing of theory with data.

Thus, we need combined educational and scientific efforts. Both must be multifaceted and coordinated. Measurement in employment settings cannot be improved without communications and education, on one hand, and scientific progress on the other.

CONCEPTUALIZATION OF VALIDITY

Validity, like Gaul, has been conceptually divided in three parts, since the publication in 1966 of *Standards for Educational and Psychological Tests and Manuals* (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1966). The division of validity into criterion-related, content, and construct parts has become standard in psychology. This conceptualization, however promising it may have appeared in the days before many of the practical issues of current concern had emerged, does not serve as well today. Possibly, the tripartite division of validity has had more relevance for educational and clinical settings than it has for employment situations.

Also, many persons and organizations (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice) apparently have considered this division of validity more concrete than its framers intended. For example, *Standards for Educational & Psychological Tests* (American Psychological Association, 1974) spoke of criterion-related, content, and construct as "aspects" of validity and stresses their logical and operational interrelatedness. Certainly nothing in this document appears to warrant the stance that the government agencies (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice, 1978) have taken, which so categorically applies different rules of evidence for criterion-related, content, and construct validity.

Various authors have taken issue with the rigid categorization of validity (Cronbach, 1980a, 1980b; Dunnette & Borman, 1979; Guion, 1977, 1978, 1980; Messick, 1975, 1980; Tenopyr, 1977; Tenopyr & Oeltjen, 1982). Moreover, in its statement of standards for selection procedures, the American Psychological Association, Division of Industrial and Organizational Psychology (1980) spoke in terms of *strategies* of validation and pointed out that the three traditional aspects of validity are really inseparable and do not necessarily represent differences in concept.

It appears that some conceptualization at a finer level than one major overall idea of validity is necessary to provide guidance for practitioners; the tripartite division does not appear to work well. Yet, at the same time, one must recognize that much of what has been said under the rubric of the three-category system has value and should not be lost.

Few would disagree that all validity is essential construct validity (Anastasi, 1976; Cronbach, 1980a; Guion, 1980; Loevinger, 1965; Messick, 1975, 1980; Tenopyr, 1977; Tenopyr & Oeltjen, 1982). However, what is needed is a conceptual framework to guide one to achieving construct validity. In developing such a framework, the following considerations are expounded upon as they relate to employment testing:

(a) validation strategies are largely situationally determined with the investigator's specific purpose being paramount,

(b) validity can be conceptualized along a continuum from *specific to general* without the imposition of rigid categories of validity,

(c) *content validity* is essentially a meaningless term,

(d) criterion-related and content-oriented strategies are closely interrelated, are only strategies and means to an end of contrust validity, and, depending upon the exact circumstances of test development and research, can fit at various points on the continuum from specific to general validity.

Situational Determination

For employee selection, in particular, it appears that the validation strategy, which will be optimal, is to a large extent situationally determined. It has long been held that it is the validity of inferences from test scores about which we should be concerned (Cronbach, 1971). One of the major problems in employment settings, is that such inferences usually must be made in a dynamic situation, whereas the typical modes of test validation to a great extent assume a static situation. For example, when one embarks upon a criterion-related study, one gets a criterion at a particular point in time. Tradition holds that the criterion must be maximally relevant for conditions that exist at that point in time. For example, if a criterion is a measure of job performance, the job duties involved in criterion measurement must be those which are actually done at that point in time. If, however, the job changes, as most jobs do, the criterion may no longer be relevant, and the validation study results and inferences based thereon will be, at the best, considered ambiguous. Either criteria must be broadened so that they become more general, such as substituting supervisor's ratings for work samples, or new validation studies must be done to accommodate ever changing criteria.

In a typical employment situation, jobs do not remain constant; the notion of a fixed job simply must be dismissed. One of the things personnel selection psychologists have to cope with is the ever-changing job. Sometimes it is found that in a long predictive study, the job involved changes so that the early subjects are not doing the same job as later subjects. Furthermore, job context factors are often changing. Although it is not likely, some of these may serve to alter validation study applicability. Applicant populations also change; although many applicant characteristics do not affect validation results (National Research Council, 1982), there may be some that do. Finally, in any employing organization, jobs must be grouped in some way for admistrative purposes. For example, most employers would not change a secretary's pay or cause him or her to be retested when moving from one supervisor to another, regardless of the differences in styles of supervisors and the ways they utilize their secretaries. In any validation effort, jobs also must be grouped. It is seldom one in practice encoun-

ters a situation in which everyone in a validation sample does exactly the same job. If one strictly followed typical validation tenets, one might be able to muster at the most *N's* of two or three. The situation with job grouping has many of the same effects as that with job changes. Narrow, job-specific criteria will not usually result in validation results that support the inferences one needs to make. What has been said of criteria also applies to predictors developed on the basis of content or psychological theory.

Specific and General Validity

It appears that, at least for employment settings, there must be some reconceptualization of validation. Other authors (Cronbach, 1971; Loevinger, 1957) have pointed out the *ad hoc* nature of most validation efforts and the need to extrapolate in all validation whether in an employment setting or not. The limits of permitted extrapolation depend on how one developed one's validity evidence in the first place. No precise rules for extrapolation can probably ever be developed, but some new ways of thinking about validity, which may aid in making judgments about inferences from tests or other measuring devices, appear to be in order.

It is proposed that there is a continuum on which, at one end, is *specific validity* and, at the other end, *general validity*. Neither of these two terms signifies a type or component of validity. They just represent extremes differentiated only by a shift in emphasis. Most validation results will fall somewhere between the two extremes. In many ways, the two terms denote many of the conditions Cook and Campbell (1979) described when they spoke of *internal* and *external* validity. The term *specific* roughly corresponds to the term *internal,* and *general* is close in meaning to *external.* The new terms have been chosen because the meanings do not exactly coincide with those of the older terms, and confusion with the teachings in experimental psychology might result were different terms not used.

Specific validity occurs when one designs a study so that the results will have a high fidelity in a given situation, in a given location, for a specific population, at a specific point in time. If one does his or her work well, inferences within the confines of the given situation will be relatively accurate. Yet, if the situation is at all dynamic and/or generalization to a similar situation is required, one has little evidence upon which to proceed. An example would be a job knowledge test for machinists, which would not be so applicable to stock clerks.

General validity occurs when one designs a study so that results will have generality for a number of situations. Usually, it can be expected that the inferences relative to any one situation in the set of situations covered will not be so accurate as they would be had the study been done using procedures more appropriate to the *specific* end of the continuum. An example would be a verbal

aptitude test which could be expected to have some validity for both machinists and stock clerks.

A general hypothesis can be stated regarding *specific* and *general* validity. That is, both cannot be maximized at the same time. In general, to increase one is to decrease the other. As one moves away from the specific end of the continuum, one automatically moves toward the general end and vice versa. Ultimately, the continuum of test development depends highly upon one's purpose and the exact situation.

It is difficult to test this hypothesis, as most organizations will not support the type of research involved. For example, the typical development of highly specific work sample tests, e.g., data-entry tests, involves a situation in which tests and any appropriate cirteria are so similar that a criterion-related validation study would result in a validity coefficient which would approximate a test-retest reliability coefficient, e.g., (Tenopyr & Caire, 1966). Supporting content-oriented test construction is usually the only investment an organization will make in a situation of this sort. Also, an organization would not normally support efforts to show that a data-entry test is more valid for predicting data-entry performance than sheet-metal work performance. On the other hand, organizations will support the typical research that is reported in the literature, i.e., studies involving the same more general tests (aptitude tests) for a variety of jobs. It is also significant to note that even after adjustment for restriction in range and unreliability of the criteria (Schmidt & Hunter, 1977), predictive validities of these more general tests fall far short of their reliabilities.

A logical parallel may be drawn in the field of education. Despite the fact that it is known that a general scholastic aptitude test is a fair predictor of grades in many courses, it is a rare educator who would consider this general test to be more valid for assessing classroom performance than a specific test requiring mastery of what was taught in the class. Nor would an educator conduct research to determine whether a classroom algebra mastery test was as valid as a classroom English composition test in predicting performance in composition.

Because a research base will probably never be developed to determine the tenability of the hypothesis outlined, the notion of the incompatibility of specific and general validity will probably never achieve more than the status of a working hypothesis.

Also, it should be noted that the notion of specific v. general validity applies most logically in the context of predicting performance; whether it would apply in situations where criteria like tenure are predicted is a research question.

Specific Validity

An example near the specific end of the continuum would be a work simulation which had a high fidelity to the duties of a specific position. If a screw is to be turned to the left on the job, it is turned to the left in the simulation. However, the

notion that one can exactly duplicate the work in a testing situation is a fiction. Every test is an abstraction. Some tests are just less abstract than others. The least abstract are probably the flight simulators, whose technology is far too expensive to duplicate in normal employment situations. Even the supposedly simple typing test is an abstraction. In fact, the typing test presents a good vehicle to demonstrate the necessity of abstraction in employment testing. First, there is the question of the material to be typed. In an organization of any size, one will find wide inter- and intraindividual differences in many characteristics of the material typed. For example, one person types only one- or two-paragraph memoranda. Another types a combination of memoranda and statistical reports. Some production typists may encounter all types of work. The work for an individual typist may vary from day to day. In developing material for a typing test, one is faced with a number of dilemmas. However carefully one samples the material typed in an organization, the resultant material selected for the actual test or tests will be a compromise of some sort and probably not reflect what any given typist in the organization actually types on a given day. Considerations relative to the job applicant population must also be taken into account. for example, in an engineering firm, does one include in the test technical words that a person in a high school typing course probably has never encountered? There are other considerations. If it is found that most typists type from handwritten copy or edited drafts, whose style of penmanship does one use for the test material? How clear and consistent should the editing be—like a professional editor's work or like the chicken scratches of a harried manager? Is spelling to be corrected? Are the length of the test and time limits to reflect the duties of a busy secretary who cannot type for more than 5 minutes without being interrupted, or the activities of a word processing production typist who is expected to type over long periods? How should speed and accuracy be weighted? In view of the employing organization's policies on job classification, pay, and employee mobility, can more than one test or a test with different critical scores be used?

Equipment and job applicant-equipment interactions must be considered. With all of the varieties of typewriters and word processors available today and often coexisting in a given employing organization, equipment choice is very difficult. Furthermore, one must consider that many applicants may not have had training on any of the equipment used in the organization, and one may wish to measure basic skills as opposed to equipment-specific skills.

Also, equipment considerations interact with content choice. For example, if hyphens are at different places on various keyboards, one may wish to eliminate typing of hyphens from the test content. Consequently, equipment considerations may serve to add to the abstract nature of the test, making it far from an actual job sample.

Perhaps the highest specific validity, at least in concept, is achieved by well designed probationary periods or documented experience in the work involved.

Also, in concept, it would be expected that lower internal validity would be associated with aptitude tests or general education requirements.

Criterion-related strategies may fall anywhere between the extreme of the two ends of the specific-general continuum. The exact placement depends on the nature of both the predictors and the criteria. If one uses as a predictor a very specific test, designed for the particular job, and employs a criterion which accurately reflects specific job requirements, one's validity will probably be nearer the specific than the general end of the continuum. Various combinations of specific and general predictors and criteria can exist; consequently, one has to examine the exact situation to estimate how general or specific one's criterion-related validity is.

Experienced researchers recognize that specific validity is not necessarily optimal, despite its intuitive appeal. For example, the more faithful a replica of a job a work sample is, the more likely it is to have to be changed constantly to accommodate changes in detailed job procedures. If, perchance, performance on one's detailed work sample involves constructs that have broad generality, one should have additional evidence to defend generalization. Furthermore, in employment settings, face validity takes on importance with the psychologist's clients. For example, a test battery for telephone operators once contained a test involving completing mark-sense cards. The job of telephone operator was changed to eliminate the use of such cards. Thereafter, the supervisors of telephone operators assumed that the whole test battery was not useful in selecting operators, despite the fact that the test was still valid. A more practical strategy might have been not to strive for less specificity in predictors.

Content Validity and Specific Validity

As every test, even the supposedly simple typing test is an abstraction; the very notion of content validity is called into question. Content sampling for the purpose of selection-device construction always results in something other than a job replica. The specific end of the continuum may be more easily approached in educational achievement testing, where sampling from what is taught is a somewhat simpler task than sampling in a dynamic job situation. However, even in educational testing, it is probable that true specificity is seldom achieved.

Content validity as a concept has been criticized for a variety of reasons (Guion, 1977, 1978; Messick, 1975; Tenopyr, 1977). Messick (1975), in particular, has proposed that what is typically called content validity is concerned with inferences about test construction, not individuals. Tenopyr (1977) has proposed that content only be considered one form of evidence for construct validity.

Nevertheless, in employment settings, content cannot be ignored in trying to achieve specific validity. If one wants a high fidelity selection procedure, even though it may have little generality, content-oriented strategies in test or criterion

development must be used. However, one must always remember that the inferences one makes are on the basis of constructs, however narrow they might be. A limited concept like the "ability to type numerals" is indeed a construct. If one wants to infer constructs, not made obvious by the content of the test, other evidence such as results of a criterion-related study must be brought to bear.

It should be noted, however, that some tests developed for specific, narrow purposes may have more generality than is apparent. For example, performance in a drafting test may be related to performance in a drill press operator's job. This generality may be artifactual, e.g., both draftspersons and drill press operators are trained in the specifics of blueprint reading. However, there may be some commonality of more basic constructs between a draftsperson's and a drill press operator's job requirements. Space visualization is a likely candidate. Again, evidence other than content that generalization is possible should be developed.

This is not to say that content alone cannot be the only evidence of validity. There are many situations in which content-oriented evidence of validity is sufficient, despite the difficulties in moving from inferences about test construction to inferences about individuals. Most of these, however, will be toward the specific end of the continuum. Certainly the more general interpretations should be supported by more than content. No precise rules can or should be formulated to fit all situations. Whether one chooses to use content considerations alone requires the exercise of professional judgment, taking all situational factors into account.

General Validity

General validity refers to the end of the continuum where the inferences to be made are less situation-specific. At the specific end of the continuum, one might make inferences about the ability to enter numeric information in a computer terminal. At the general end, one's inferences would reflect abilities more like that to do general clerical work.

These inferences differ mainly in their specificity. They do not differ in kind. Both reflect constructs; the more specific inference reflects a narrow construct, presumably largely supported by a wide variety of evidence, which may include results of a criterion-related study and does not necessarily exclude content. However, when one is attempting to support a general inference, it appears that there would be few situations in which content alone would be sufficient evidence.

As mentioned previously, in most employment situations, it is the more general validity in which one is interested. One normally needs to make inferences about behavior in more than a narrow band of situations. How much validity can be extended to a variety of situations is a matter which has been discussed in the courts (*Douglas v. Hampton* et al., 1975). Pearlman (1980) has

indicated how job grouping can be done on the basis of test validity. If employers did indeed usually group jobs on the basis of ability-related job requirements, the psychologists' situation relative to marshaling evidence of general validity would be much simpler. However, most employers do not have professional psychologists doing job analysis, job evaluation or job grouping for progression and pay purposes. These matters also are often bargained for, making them even farther from the psychologists' control. Also, there may be wide intercompany differences and, even within the same company, interdepartmental differences. What is considered a job in one company or department may be considered a group of several jobs in another department or company.

In addition, where systems of job grouping and progression are developed, in some companies it has not been uncommon for such systems to reflect biases of various sorts. For example, jobs normally populated by persons of one sex are grouped together, regardless of differences in skill and ability requirements.

Personnel psychologists find it easier to work within existing systems than to try to change them. The ethical dilemmas involved are not discussed here, but needless to say, they are many.

Working within these existing job systems, psychologists probably still can do much to effect valid selection procedures. The question of whether a given predictor has generality enough to be used for groups of jobs is largely, although not entirely, an empirical question. It is not feasible to attack the problem wholly by strict empiricism. For example, how much of a job change or difference in jobs dictates a new validation study is a judgment question. How much lowering of validity in the specific situations one is willing to tolerate when using a general predictor is also a matter of judgment.

Although empiricism can take many forms (Cronbach, 1971), it can consist of criterion-related studies relative to a sampling of jobs in the job group in question. Validity generalization then can be helpful in extrapolating to jobs not in the sample.

Validity generalization to date has been discussed in terms of which tests are valid for which jobs (Callender & Osbourn, 1980; Schmidt, Gast-Rosenberg, Hunter, 1980; Schmidt & Hunter, 1977; Schmidt, Hunter, Pearlman, & Shane, 1979). Another perspective on validity generalization is taken here. It is suggested that validity generalization research tells us as much about criteria as it does about tests. In discussing any relationship such as those indicated by coefficients of correlation, both variables underlying the relationship must be considered.

In particular, both predictors and criteria must be considered relative to their generality. Most of the validity generalization research has been done on aptitude tests which fall near the general end of the continuum. The criteria employed in these studies have been, for the most part, supervisors' ratings. These also are highly general. There has been little research involving either more specific tests or criteria.

The generally positive validity generalization results obtained to date, in the author's opinion, represent essentially the operation of only those abilities supervisors discern most readily in a work situation. The halo effect attendant upon supervisors' ratings is well known, and it appears that typical supervisors' ratings reflect only the grossest of behavior. Were one's criteria able to capture the nuances of behavior in given jobs, one might have a better basis for inferences about generality. This is not an easy task. Anyone who has attempted to develop tests to measure specific abilities knows well the large number of false starts associated with this effort. Test tasks which, on the surface, appear to measure the same ability many times, indeed do not. Other tests designed to measure different abilities instead measure the same ability. Except at the most rudimentary level, the endeavor to glean ability requirements from job duties is even more difficult. Experienced investigators doing validation research know well that, despite the results of job analyses, some "shotgunning" of predictors is still a viable research strategy. The problem is compounded with jobs in which the manner in which one performs the job is to any extent discretionary, and different abilities may be used by different persons to achieve the same performance levels. Also, improvement of prediction of behavior in employment settings is much needed. As Ghiselli (1966) pointed out, prediction of job performance has not been highly impressive. If we are to improve prediction, we must design any validity generalization research carefully. We should pay as much attention to the criterion-side as we pay to the predictor-side. It is the author's opinion that the most meaningful validity generalization research would be that in which criteria are relatively specific so that abilities that might be obscured by a more general criterion can be captured.

If we design our research this way, we may achieve a more optimal point on the specificity-generality scale than has been achieved by validity generalization research to date. By dealing with both general predictors and general criteria, this research has indicated that we can achieve moderate prediction of performance in a wide range of jobs with a few general types of tests. If we are to improve prediction, we are probably going to have to move toward the specific end of the scale in terms of both criteria and predictors and be satisfied with less generality. Again, how far one moves on the continuum is a matter of judgment and, to a large extent, influenced by situational factors.

Research of this sort would also enable one to do a better job of developing taxonomies. In this respect, it should be pointed out that the basis for job taxonomies also form a continuum from specific to general. Job taxonomies can be formed on anything from a narrow to a broad basis. For example, a job which involves turning a screw to the left instead of to the right as in another job may be put in a different family from the other job. At the other extreme, approaching jobs from a worker attribute rather than a task approach and assuming that all jobs involve some overall ability, all jobs could be grouped in one family. The level of generality one chooses as a basis for taxonomies and where one estab-

lishes taxonomic boundaries should, as in testing, have some empirical support, but are in the end judgment calls.

Content, Criterion-Related and Construct Strategies

Two of the three traditional strategies of validation, content and criterion-related, have largely been presented here as means to an end. That end is construct validity. The necessity of the use of professional judgment in determining which strategy or which combination of strategies one employs has been emphasized. The role of situational factors has been indicated to be important and a major basis for judgment.

The type of strategy one uses and, consequently, the evidence of validity one amasses cannot be dictated by precise rules. As has been indicated, content can be a form of evidence anywhere along the continuum from specific to general. However, it becomes the major form of evidence near the specific end of the scale. Criterion-related strategies cannot be divorced from content strategies. Content is usually a major consideration in criteria. Criterion-related strategies can form evidence anywhere along the continuum depending upon the generality of the predictors and the criterion.

Construct validity, which should be the basis of all inferences from psychological measurement, of course, cannot be separated from the strategies used to achieve it. Construct validation can draw from a number of lines of research and is not a simple matter (Cronbach, 1971). Defining precise measurement steps for achieving construct validity as some have done (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice, 1978) tends to belie the complexity of all measurement situations and freeze the state of science at a rudimentary level.

Construct Interpretation

One of the greatest problems in industrial and organization psychology today is the inability of practitioners to make construct interpretations from their measurements. This is not a problem solely for this group of psychologists, but is a difficulty throughout psychology. Witness the number of tests in print (Buros, 1978). Were there available meaningful systems of constructs or, for that matter, any systematic efforts to establish construct validity except for a few major tests, there would be less test development. If test reviewers had meaningful construct relevant information with which to evaluate new tests, perhaps test authors would be less enthusiastic in making claims of having measured something new and different. Certainly, the "made-up-on-the-spot" construct has led to considerable confusion.

General validity could more easily be achieved were there the possibility for more well supported construct interpretations. If systems of constructs were

available, the need for *ad hoc* evidence of general validity in each validation effort would be considerably lessened.

Science, of course, cannot progress without constructs; in fact, well developed systems of constructs are the mark of a fully developed science. If psychology and employment psychology, in particular, are to mature, they must do more to achieve the bases for construct interpretation of data. Meaning must be added to measurement. Numerous *ad hoc* studies, as has been the tradition of employment psychology, can result in the evolution of some principles, but seldom do they result in the explanations that are so much needed.

In the ability area, psychologists have the work of Ekstrom, French, and Harman (1979) to which to turn. This monograph covered well the status of aptitude constructs at its date of completion and can be used to support the construct interpretations of various types of aptitude test. The work can also be used judiciously in establishing general validity.

Unfortunately, in areas such as personality or character measurement, there appears to be no counterpart work to which to turn. In attempting personality measurement, one is faced with numerous unsupported and often contradictory claims of construct validity. Because of the various problems attendant upon personality measurement, e.g., invasion of privacy, low validity, inventories in this area are not used much by employers (Bureau of National Affairs, 1983). However, if some of the problems of construct confusion were eliminated, there might be some possibility that predictors in the noncognitive areas could become more useful in employment testing than they now appear to be.

Another area in which construct systems are needed is organizational psychology. This field is now characterized by a myriad of questionnaires and rating scales of various sorts, which are purported to measure things like job satisfaction and job commitment. These are part of the whole employment process and may be conceived of as potential criteria in certain situations. Such questionnaires and scales have most of the problems associated with personality measurement.

If we are to have general validity, not just in the limited sense of aptitude test validation, we must get better bases for construct interpretations through the whole range of measurement techniques. It has been said that no two psychologists could agree on systems of constructs, but independent investigators, free from the constraints of any one laboratory, or other organizational setting should attempt to bring more meaning to our measurements.

A related problem, as discussed by Pearlman (1980), is the need for meaningful taxonomies of work. Unless the bases for criteria can be organized in some meaningful fashion, there is little hope for achievement of the more general validity we usually need. Attempts to obtain an all-encompassing taxonomy of work probably would be disappointing. However, it is suggested that progress toward such a taxonomy can be expedited by first attempting to develop better taxonomies of human characteristics through traditional measurement. Test

tasks, although they are necessarily limited in scope, can be the basis for systems of constructs which can be used to classify job tasks.

JOB ANALYSIS

There are many ways of analyzing jobs, depending upon one's purpose. Many of the better developed of these techniques have been reviewed by Pearlman (1980). However, it appears that practically every investigator uses somewhat different methods in analyzing jobs.

The fact that different job analysis techniques are needed for different purposes had led to some extent to the proliferation of these job analysis procedures. Also, considering that every job analytic situation is different, involving different jobs and different populations, the comparison of job analytical techniques from different investigations is made difficult. Few investigations involving use of different job analyses in the same situation seem to have been made; there are, however, some exceptions (Ghiselli, 1966).

Although a universal job analysis system is not advocated, it appears that there is a need for developing some principles for analyzing jobs. Despite the large number of job analyses which are being done today, there does not appear to be available the research base from which the needed principles can be drawn. The lack of principles to guide methodology, of course, hinders the development of the taxonomies relative to worker attributes and job requirements.

The major question of the validity of the masses of data which have been generated is of utmost importance. Pearlman (1980) has suggested that test validities and the results of validity generalization studies be used to form job groups based upon abilities required of the incumbents. Unfortunately, there are some problems associated with this approach. In particular, the job groupings afforded may be too broad to use for a particular purpose and may not reflect the more specific ability requirements in different jobs.

Approaches involving having supervisors or job incumbents rate jobs on construct-oriented scales were advocated by this author (Tenopyr, 1977). These methods do not seem so appealing upon reconsideration. The main problem is that there is a dearth of evidence that job experts can rate jobs validly in terms of their ability requirements. The often demonstrated finding that job experts can agree on ability constructs needed for job performance, of course, supports reliability for such ratings, but there seems to be no evidence that these ratings are related to validities of corresponding tests.

What is needed is a series of studies which attempt to determine validity of construct estimates by job experts. Various types of rater should be examined, e.g., supervisors, incumbents, psychologists. Different specificity levels of construct should be employed. Studies to determine the degree of response style associated with such ratings should be undertaken.

JOB PROFICIENCY MEASUREMENT

Methods for measuring job proficiency or performance have been a subject of study for many years. The studies in this area have been reviewed by Tenopyr and Oeltjen (1982). They represent special measurement problems of their own but become additionally problematic when used as criteria in validation.

There are essentially two categories of measurement involved in evaluating proficiency, the supervisors' rating and the objective record of performance. Some of the problems with supervisors' ratings have been discussed previously. Supervisors' ratings developed for the organization's administrative purposes pose special problems. The most important of these is the coupling problem. Rating results are often coupled with administrative actions such as salary increases, promotions, or personal development counseling. Rumors abound in every organization of supervisors who "back into" a rating, e.g., they decide on the amount of the raise first and then give a rating to justify it. When ratings are tightly coupled with one administrative purpose, they are often found useless for other purposes to which they are less tightly coupled. Unless operational ratings are tightly coupled to all the administrative purposes for which they will be used, including feedback to the employee, or are not tightly coupled with any administrative system, they will not be maximally useful as criteria. Most practicing personnel psychologists, therefore, appear to prefer not to use in-place rating systems as a basis for criteria. They instead rely on specially developed criteria for the study involved.

Objective records, despite their intuitive appeal, have many drawbacks. For example, for welders, error rate per inches of weld made might be considered for performance measurement. However, the most proficient welders might get the most difficult welding jobs, such as welding corners of boxes or joining materials that are difficult to weld. In a factory situation, the most senior, but not necessarily the most proficient, operator may get the newest and most efficient machine.

A major problem is that in place performance measurement systems are often gamed. A plant manager may turn out high, short-term profits by skimping on maintenance of the factory. His or her successor may then have to do the maintenance and, thereby, turn in poorer profit picture. The phenomenon of employees' paying attention to those phases of the job upon which pay and promotion are based and neglecting other job aspects is common. Unfortunately, those who have tried to develop operational performance measurement systems have generally found it impossible to cover all aspects of any job and thus reduce the possibility of "gaming."

Another problem with objective performance measurement is that of getting a large enough number of observations to get reliable measurement. This is particularly true when error rate is small. Also the task of obtaining and summarizing

the data is often so administratively burdensome that employers avoid sophisticated performance measurement systems.

A final problem is that operational performance ratings are often not available in unionized operations where personnel decisions are made largely on the basis of seniority.

Thus, when personnel researchers want objective criteria, they are often forced to develop *ad hoc* measures for any study involved. Even then, there are often administrative difficulties in getting supervisors to make enough systematic observations to obtain reliable measurement.

Despite the serious problems in this area, research on performance measurement should continue. As more and more jobs are involving automated equipment, the probability of sufficient, accurate data in simpler jobs is increased. Also, larger computer-based measurement systems are being made possible. The many problems with supervisors' ratings will not be solved easily. Probably researchers, if forced to use supervisors' ratings as criteria, will continue to develop them on an *ad hoc* basis.

ALTERNATIVES TO TESTS

One of the needs for new direction for measurement in employment settings is to provide better development guidance for the alternatives to paper-pencil tests. In particular, the employment interview, which is in wide use (Bureau of National Affairs, 1983), needs further development (Reilly & Chao, 1982; Tenopyr & Oeltjen, 1982).

Much has been published about what goes on in the interview. Tenopyr and Oeltjen (1982) found that over a recent 3-year period, there were sixteen studies involving the effects of race and sex upon interview results. Most of these were of the ''paper people''-type which involved identical descriptions of people, except for race or sex. Unfortunately, in this same review, only one validation study for an interview was found.

A dynamic situation like an interview is not an easy subject for study. Often when research is done, it is necessary to reduce this fluid situation to written form. Paper people is one vehicle; casting interview questions so that they are nothing more than an orally administered biodata blank is another.

It appears that much is known about the pitfalls of interviewing; now is the time to work on the development of valid interviews. It is a much easier task to examine minutely existing practices than to develop new practices which actually work.

Other widely used procedures for which there is little guidance for developmental practices are experience evaluation methods and work samples. The former are of as much importance as the interview because they are so widely

used. Certainly their validity to date has not been impressive (Caplan & Schmidt, 1977).

DIFFERENTIAL PREDICTION

Possibly no discussion of new directions in measurement can ignore the question of group differences in regression systems. The research in this areas, as reviewed by Tenopyr and Oeltjen (1982) has been abundant, if controversial. This whole line of research has taken a course which is perplexing to a scientist. The long search for differential validity and differential prediction systems has taken place without any reasonable scientific hypotheses as to why such phenomena should be found. Unless science can be incorporated into this research and meaningful hypotheses generated, it is suggested that such research receive less emphasis. Any further research, such as that into sex differences in regression systems, should be more carefully designed so that artifacts, such as differences in exposure to certain kinds of training, do not lead to erroneous interpretations of results. Investigators should also be certain that criteria have the same meaning for all groups concerned. Too often, factors such as affirmative action programs or attitudes of supervisors and trainers may render ratings or, even training grades, unsuitable criteria. Certainly, in the absence of meaningful scientific hypotheses, researchers bear a heavy burden to prove any group differences found are not artifactual.

SUMMARY

Measurement in employment settings is fraught with many difficulties, some of which are unique to personnel selection. Unless the organizational support for sound measurement is obtained, these difficulties will be with us for many years to come. A synergistic combination of education for organizational personnel and application of science to measurement in organizations is needed.

A more flexible reconceptualization of validity coupled with a renewed emphasis on interpretation of data in terms of constructs is required. Methodology needs to be improved to achieve these ends. There needs to be a revitalized effort to improve interviews and other techniques which are far more widely used than paper-pencil tests. Finally, science should be incorporated into research on group differences and, unless rational scientific hypotheses can be generated and tested relative to differential prediction, such research should receive less emphasis.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, D.C.: American Psychological Association.

American Psychological Association, Division of Industrial-Organizational Psychology. (1980). *Principles for the validation and use of personnel selection procedures.* Berkeley, CA: Author.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Bureau of National Affairs. (1983). *ASPA-BNA survey no. 45. Employee selection.* Washington, D.C.: Author.

Buros, O. K. (Ed.). (1978). *The eighth mental measurements yearbook.* Highland Park, NJ: Gryphon.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology, 65,* 543-558.

Caplan, J. R., & Schmidt, F. L. (1977). *The validity of education and experience ratings.* Paper presented at Annual Meeting of the International Personnel Management Association. Assessment Council. Kansas City, Missouri.

Cook, T. D., & Campbell, D. T. (1979). *Quasi experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1980a). Selection theory for a political world. *Public Personnel Management, 9,* 37-50.

Cronbach, L. J. (1980b). Validity on parole: How can we go straight? *New Directions in Testing and Measurement, 5,* 99-108.

Dennis, E. F. (1979). *Accounting for slower economic growth—The United States in the 1970's.* Washington, D.C.: The Brookings Institution.

Douglas v. Hampton et al. (1975). 512 F2d. 976.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification. In M. R., Rosenzweig, & L. W. Porter, (Eds.), *Annual review of psychology, 30,* 477-525.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monograph,* No. 79-2.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal register, 43,* 38290-38310.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Griggs v. Duke Power Co., 401 U.S. 424 (1971). 3EPD P8137, 3FEP175.

Guion, R. M. (1977). Content validity, the source of my discontent. *Applied Psychological Measurement, 1,* 1-10.

Guion, R. M. (1978). Content validity in moderation. *Personnel Psychology, 31,* 205-214.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11,* 385-398.

Hoffman, B. (1962). *The tyranny of testing.* New York: Crowell-Collier.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72,* 143-155.

Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955-966.

Messick, S. A. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

National Research Council. (1982). *Ability testing: Uses, consequences, and controversies.* A. K. Wigdor & W. R. Garner (Eds.), Washington, D.C.: National Academy Press.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin, 87,* 1-28.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1-62.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643-661.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology, 32,* 257-281.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30* 47-54.

Tenopyr, M. L., & Caire, J. (1966). *The validity of a key punch test.* Unpublished research report. El Segundo, California: Rockwell International.

Tenopyr, M. L., & Oeltjen, P. D. (1982). Personnel selection and classification. In M. R., Rosenzweig, & L. W. Porter, (Eds.), *Annual review of Psychology, 33,* 581-618.

Whyte, W. H. Jr. (1956). *The organization man.* New York: Simon and Schuster.