DOCUMENT RESUME

ED 348 375		TM 017 305

TITLE	Proceedings of the 1989 IPMAAC Conference on Personnel Assessment (13th, Orlando, Florida, June 18-22, 1989).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	Jun 89
NOTE	254p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	Assessment Centers (Personnel); Computer Assisted Testing; Computer Simulation; Decision Making; *Evaluation Methods; Job Analysis; *Job Performance; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; *Public Sector; Test Use
IDENTIFIERS	International Personnel Management Association

ABSTRACT

The International Personnel Management Association Assessment Council (IPMAAC) is a non-profit organization of personnel assessment professionals involved in public personnel assessment. Author-generated summaries/outlines of papers are presented. The keynote address is "Ability Testing in the 1980's and Beyond: Some Major Trends" by A. Anastasi. The categories of the summaries of 52 papers include: management issues and innovations; professional and legal issues; assessment center issues; job analysis techniques and research; test construction and validation; uniformed forces testing; bio-data uses; screening and retaining applicants; and developing physical ability standards. Selected topics within the categories include: use of a job element questionnaire and interview to select public safety communications specialists; assessing creativity in a content-valid test; observations of an expert witness; agency-rated scores as an alternative to traditional education and experience ratings for highly specialized job classes; development of valid computer-based tests for assessing divided attention in nuclear power plant operators; strategies for making cut-score determinations on a performance-based observational test; the Missouri School Superintendent Assessment Center Certification Program; empirical validation of firefighter vision standards; the development of tailored response testing; criterion validation of a preemployment psychological test for correctional officers; decision making in assessment centers; assessment centers and bio-data--approaches to managerial selection; video-based structured interviews and testing; managerial incompetence; the development of physical ability standards; the impact of physical standards projects on internal race and sex relations; recent developments in job analysis research; oral board examinations; assessing teacher candidates' writing; the Bush Administration and the 101st Congress; a computerized tracking system; and organizational perspectives on the setting of cutting scores. An author index is included. (SLD)

IPMA Assessment Council

Proceedings of the 1989 IPMAAC Conference on Personnel Assessment

June 18-22, 1989
Orlando, Florida

PROCEEDINGS OF THE THIRTEENTH ANNUAL INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL, 1989

The PROCEEDINGS are published as a public service to encourage communication among assessment professionals about matters of mutual concern.

The Proceedings essentially summarize the presentations made at the Conference. All presenters, with the exception of invited and keynote speakers, were required to limit the published version of their papers to approximately 4 pages. Hence, some presenters were able to include most of their oral presentations, while others opted to provide only topical outlines. Papers were published in the condition received, without editing. A few authors did not provide a written version of their presentations, and hence their presentations are not included in this volume.

Persons interested in additional information regarding a presentation should contact the author(s) directly in order to determine if a more complete paper is available.

Prepared by:

Esther Juni
1989 Program Chair
New York City Department of Personnel
220 Church Street
New York, N.Y. 10013

The INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL (IPMAAC) is a non-profit organization of personnel assessment professionals. IPMAAC's members are actively engaged in, or contributing to, professional level public personnel assessment.

IPMAAC was formed in October 1976 to provide an organization that would fully meet the unique needs of public sector assessment professionals by:

- providing opportunities for professional development;
- defining appropriate assessment standards and methodology;
- increasing the involvement of assessment specialists in determining professional standards and practices;
- improving practices to assure equal employment opportunity;
- coordinating assessment improvement efforts.

IPMAAC OBJECTIVES support the general objectives of the International Personnel Management Association—United States. IMPAAC encourages and gives direction to public personnel assessment; improves efforts in fields such as, but not limited to, selection, performance evaluation, training, and organization effectiveness, defines professional standards for public personnel assessment, and represents public policy relating to public personnel assessment practices.

IPMAAC EXECUTIVE COMMITTEE
Joel P. Wiesen, President
Sally McAttee, President-Elect
Nancy Abrams, Past President

TABLE OF CONTENTS

KEYNOTE ADDRESS

MANAGEMENT ISSUES AND INNOVATIONS

PROFESSIONAL AND LEGAL ISSUES

ASSESSMENT CENTER ISSUES

MANAGEMENT ISSUES AND INNOVATIONS

JOB ANALYSIS TECHNIQUES AND RESEARCH

TEST CONSTRUCTION AND VALIDATION

UNIFORMED FORCES TESTING

ASSESSMENT CENTER ISSUES

PROFESSIONAL AND LEGAL ISSUES

TEST CONSTRUCTION AND VALIDATION

Page

ASSESSMENT CENTER ISSUES

TEST CONSTRUCTION AND VALIDATION

UNIFORMED FORCES TESTING

TEST CONSTRUCTION AND VALIDATION

UNIFORMED FORCES TESTING

ASSESSMENT CENTER ISSUES

BIO-DATA USES

TEST CONSTRUCTION AND VALIDATION

SCREENING AND RETAINING APPLICANTS

WRIPAC INVITED SPEAKER

PRESIDENTIAL ADDRESS

TEST CONSTRUCTION AND VALIDATION

DEVELOPING PHYSICAL ABILITY STANDARDS

JOB ANALYSIS TECHNIQUES AND RESEARCH

TEST CONSTRUCTION AND VALIDATION

TEST CONSTRUCTION AND VALIDATION

TEST CONSTRUCTION AND VALIDATION

PROFESSIONAL AND LEGAL ISSUES

SCREENING AND TRACKING APPLICANTS

TEST CONSTRUCTION AND VALIDATION

ABILITY TESING IN THE 1980s AND BEYOND: SOME MAJOR TRENDS

Anne Anastasi

Fordham University

As I look at what is happening in testing today, I am impressed by both the magnitude of the changes and the rapidity of their development. In contrast to earlier decades, the pace has increased spectacularly, and the changes tend to be basic rather than superficial. For both reasons, it is likely that the trends we see emerging today provide a preview of what testing will be like in the decades ahead.

In my efforts to touch on the highlights of these changes, I have chosen a few developments that I consider particulary significant. These developments fall quite naturally under three major headings: the role of the test user, technical methodology of test construction, and substantive interpretation of test scores.

Role of the Test User

A conspicuous recent trend in mental testing is the increasing recognition of the part played by the test user. Common criticisms of testing and popular antitest reactions are often directed, not to characteristics of the tests, but to misuses of the tests in the hands of inadequately qualified users in education, industry, clinical practice, and other applied contexts. Many of these misuses stem from a desire for shortcuts, quick answers, and simple routine solutions for real-life problems. All too often, the decision-making responsibility is shifted to the test. The test user loses sight of the fact that tests are tools, serving as valuable aids to the skilled practitioner, but useless or misleading when improperly used.

The increasing focus on the responsibility of the test user is evidenced in the successive editions of the test standards published by the American Psychological Association and prepared jointly with two other associations concerned with testing (American Educational Research Association and National Council on Measurement in Education). In the successive editions published in 1954, 1966, and 1974, increasing attention was devoted to test use. The role of the test user becomes especially prominent in the latest edition (published in 1985), where it is demonstrated in several ways. The title has now been changed from *Standards for Educational and Psychological Tests* to *Standards for Educational and Psychological Testing*. The substitution of "testing" for "tests" in the later title reflects the broadened scope of the *Standards*: it calls attention to the process of test use in addition to the technical qualities of the tests themselves. The organization and content of the latest *Standards* fully support this orientation. One section (including 5 chapters) covers technical standards for test construction and evaluation; three sections (including 11 chapters) are devoted to standards for the use of tests in different professional applications and their use with special populations, as well as standards for test administration, scoring, and reporting,and for protecting the rights of test takers.

Test publishers, as well as test-related committees of national professionai associations, have also been giving increasing attention to the key role of the tect user. Some of the major test publishers are making special effotts to improve their communication with test users, to provide the necessary information for proper test selection and score interpretation, and to help guard against common misuses of tests. Test publishers are beginning to assume some responsibility for cautioning users against popular misconceptions about what tests are designed to do and what their scores mean. Attention is also being given to spelling out more fully in test manuals the necessary qualification of test users for different kinds of commercially available tests. It appears likely that in the years ahead we shall see increasing cooperative action in meeting these problems of test use, action that will involve both test publishers and national professional associations.

A noteworthy event in this direction is a recently completed project of the Joint Committee on Testing Practices,[1] sponsored by five national associations concerned with testing, including the American Psychological Association. A special working group of this committee, comprising representatives of major test publishers and association members, was the Test User Qualifications Working Group, known by the delightful acronym, TUQWoG. The chief goal of TUQWoG was to develop an empirical, data-based set of essential qualifications for users of different types of tests, which test publishers could incorporate in their test purchaser qualification forms. Over several years of intense activity, the TUQWoG project developed a most impressive data base, and some publishers have already begun to use the resulting purchaser qualification forms.

A second working group of the same joint committee has just been formed; its object is to use the available data base in developing training guidelines and training materials for test users. It is generally recognized that the surest way to improve test use and to prevent misuse is through more and better training of test users. This training may be provided at different levels and in various forms, from college and graduate courses to in-service training

[1] See Anastasi (in press); Eyde, Moreland, Robertson, Primoff, & Most (1988, in press).

programs, workshops, and continuing education activities conducted by professional associations and by test publishers, as well as by widespread dissemination of good testing practices in published sources.

The test user, as contrasted to the test constructor, is essentially anyone who has the responsibility for choosing tests or for interpreting scores and using them as one source of information in reaching practical decisions. Many test users serve both these functions. Moreover, such professional functions are in addition to the more routine activities of administering and scoring, which in many cases can be delegated to assistants who function under careful supervision. Examples of test users include teachers, counselors, educational administrators, testing coordinators, personnel workers in industry or government, clinicians who use tests as aids in their practice, and many others in an increasing number of real-life contexts. Anyone who chooses tests or uses test results needs some technical knowledge for the proper understanding of tests. If a test user, in this sense, lacks adequate background for this purpose, he or she should have ready access to a properly qualified supervisor or consultant.

The interpretation of test scores calls for knowledge about both the statistical properties of scores and the psychological characteristics of the behavior assessed by the tests. Apart from an understanding of different types of scores, such as percentiles, standard scores, and deviation IQs, the test user needs to be familiar with such basic concepts as the standard error of measurement (SEM), which serves as a corrective against the tendency to place undue reliance on a single measurement. So important is the SEM for this purpose, that the College Board now includes data on the SEM and a simple explanation of its use, not only in brochures distributed to high school and college counselors, but also in the individual score reports mailed to test takers. Even more important for test interpretation is the evaluation of score differences, as in multiscore batteries. Such evaluation requires data on the statistical significance of score differences as well as the differential validity of score patterns.

Besides correct statistical interpretation, the proper use of test scores involves an adequately informed substantive interpretation. The latter requires knowledge about the behavior domain that the tests are designed to assess, including the conditions that influence the development of relevant cognitive and affective traits. It is in this connection, too, that we hear the oft-repeated caution that a test score cannot be interpreted in isolation; it needs confirmatory data from other sources, as well as information about the individual's experiential history and about the particular contexts for which the individual is being assessed. Failure to observe this caution accounts for many current misuses of tests and for much popular mistrust of testing.

Technical Methodology of Test Construction

The increasing emphasis on qualified test users has in no way diminished the concern for psychometrically sound instruments. On the contrary, there are signs of a growing psychometric orientation in fields that began with a qualitative focus, uncontrolled procedures, and subjective interpretations. Notable examples are traditional clinical practice, behavior therapy and behavior modification, and the burgeoning field of health psychology or medical psychology (See Anastasi, 1988, chaps. 16, 19, & pp. 657-662). In all these areas, the 1970s and 1980s have witnessed an increasing recognition of the need for standardized procedures, empirically established norms, and the evaluation of such psychometric properties as reliability and validity.

Test Validation. Not only have the areas of application of psychometric procedures been expanding, but the development of testing technology itself has shown unprecedented progress in the 1980s. Such advances are especially evident in the evolving approaches to test validation (Anastasi, 1986). Some of these developments reflect trends discernible in American psychology as a whole. Notable among these trends is an increasing interest in theory and a movement away from the blind empiricism of earlier decades. This theoretical orientation is exemplified by the growing emphasis on constructs in the analysis of personality and ability, as well as in the increasing use of construct validation. The term "construct validity" was introduced into the psychometric vocabulary in the first edition of the test Standards, published in 1954 by the American Psychological Association. The discussions of construct validation that followed -- and that continue with undiminished vigor -- have served to make the implications of its procedures more explicit and to provide a systematic rationale for their use. In psychometric terminology, a construct is a theoretical concept closely akin to a trait. Constructs may be simple and narrowly defined, such as speed of walking or spelling ability; or they may be complex and broadly generalizable, such as mathematical reasoning, scholastic aptitude, neuroticism, or anxiety.

The overemphasis on purely empirical procedures during the early decades of this century arose in part as a revolt against the armchair theorizing that all too often served as the basis for so-called psychological writings of the period. But empiricism need not be blind; nor does theory need to be subjective speculation. Theory *can* be derived from an analysis of accumulated research findings and can in turn lead to the formulation of empirically testable hypotheses. Tests published since the 1970s show increasing concern with theoretical rationales throughout the test development process. A specific example of the integration of empirical and theoretical approaches s provided by the assignment of items to subtests on the basis of logical as well as statistical homogeneity. In other words, an item is retained in a scale if it had been written to meet the specifications of the construct definition of the particular scale and *also* was shown to belong in that scale by the application of statistical procedures of item analysis.

It is being recognized more and more that the development of a valid test requires *multiple procedures*, which are employed sequentially, at different stages of test construction. Validity is thus built into the test from the outset, rather than being apparently limited to the last stages of test development, as in traditional discussions of criterion-related validation in test manuals. The validation process begins with the formulation of trait or construct definitions, derived from psychological theory, prior research, or systematic observation and analyses of real-life behavior domains, such as job analyses. Test items are then prepared to fit the construct definitions. Empirical item analyses follow, with the selection of the most valid items from the initial item pools. Other appropriate internal analyses may then be carried out, including factor analyses of item clusters or subtests. The final stage includes validation and cross-validation of various scores and of interpretive score combinations through statistical analyses against external, real-life criteria.

Such a full-scale investigation in terms of multiple validation procedures is certainly appropriate for the development of new tests, and its findings should be made available in the test manual for the guidance of potential users. Nevertheless, when the test is applied in local settings, the question of demonstrating its validity for particular jobs in that setting is frequently raised. The most direct answer would seem to be a local validation of the test against specific job-performance criteria. It has become increasingly evident, however, that a properly designed criterion-related validation is usually impracticable under these conditions (Anastasi, 1988, p. 452). Among the reasons for this conclusion are (1) unavailability of sufficiently reliable and inclusive criterion data, (2) inadequate number of employees performing the same or closely similar jobs, and (3) restriction of range through preselection, since rarely can all applicants be hired and followed up. The use of current employees in lieu of applicants only introduces new problems.

In view of the practical difficulties in the way of local, criterion-related validation, several alternative procedures have been developed, and still more are being explored (for references, see Anastasi, 1986; 1988, chaps. 4 & 15). These procedures fall somewhere between purely content-related and purely construct-related validation, usually combining features of both. Some utilize refined and elaborated techniques of job analysis, as illustrated by the job element method employed by the U.S. Office of Personnel Management (Anastasi, 1986; 1988, 455-457). Job elements refer to those specific job behaviors that differentiate most clearly between marginal and superior workers. Relying ultimately on the observations and judgment of experienced workers, the job element method in its various adaptations provides descriptions of behavi ral requirements for the job. The behavioral statements can in turn be grouped into broader categories or constructs, such as computational accuracy, spatial visualization, manual dexterity, and ability to work under pressure. There is a growing body of research aimed at developing a general taxonomy of job performance in terms of relatively broad behavioral constructs. The job element method contributes to this goal and thereby facilitates the effective use of a test across many superficially dissimilar jobs.

The job element method is also related to the concept of _synthetic validation_. Insofar as job elements identify behavioral requirements common to different jobs, it should be possible to estimate the validity of a test for a particular job in the absence of local criterion-related validation. The 1980s witnessed a revival of interest in synthetic validation, which was first proposed in the 1950s (Anastasi, 1988, p. 456; Balma, 1959, p. 395). Essentially, synthetic validation involves three steps: (1) identifying the job elements and their relative weights in the job; (2) empirical analysis of each test to find the extent to which it measures each of these job elements; and (3) computing the validity of each test for the given job from the weights of these elements in the job and in the test, through a procedure that is essentially an adaptation of the multiple regression equation.

To add a very recent example of innovative approaches to local test validation, let me cite a procedure reported in the 1989 _Journal of Applied_

Psychology (Turban, Sanders, Francis, & Osborn, 1989). Designated as cognitive equivalence, this procedure was developed specifically to meet a problem of test security, which had become compromised at least in some plants of a large industrial company. The question was whether other, commercially available tests could be substituted for the current tests, by showing empirically that they measure the same constructs. The procedure involved administering both old and new tests to samples of applicants,and investigating construct equivalence through correlational analysis, structural modeling, and a comparison of the hiring decisions reached from the old and new tests.

I have mentioned these various examples mainly to show that techniques for local validation are alive and growing. It is an area worth looking into.

Strictly speaking, all practical use of tests involves some degree of validity generalization. Tests are rarely, if ever, used under conditions identical with those under which validity data were gathered. Hence some degree of generalizability is inevitably required. However when standardized aptitude tests were first correlated with performance on presumably similar jobs in industrial validation studies, the validity coefficients were found to vary widely. Such findings led to widespread pessimism regarding the generalizability of test validity across different situations. Until the mid-1970s, "situational specificity" of psychological requirements was generally regarded as a serious limitation in the usefulness of standardized tests in personnel selection. More recently, research with newly developed statistical techniques (Schmidt, Hunter & Urry, 1976) has demonstrated that much of the variance among the validity coefficients reported for industrial samples may be a statistical artifact resulting from small sample size, criterion unreliability, and restriction of range in employee samples. The subsequent accumulation of empirical evidence suggests that the validity of ability tests can be generalized far more widely across occupations than had heretofore been recognized.

It has also been shown that with an adaptation of Bayesian statistics, it is possible to combine previously available validation data on a given test with newly accumulated local validation data on small samples, thereby arriving at a combined estimate of validity that is relatively stable and generalizable (Schmidt & Hunter, 1977). The same approach can be used in combining data from different divisions or plants of a large company (Bentz, 1980, September), or by trade associations with data across member companies, as in the unusually promising consortium project of the Edison Electric Institute (Kleinke, 1987).

In general, recent, large-scale research with both industrial and government jobs has showed wide generalizability for tests of general, academic intelligence, which measure largely reasoning with verbal, numerical, and other abstract material. This is especially true for higher level professional and administrative jobs. Although other aptitudes are needed for classification decisions and for more specialized skill jobs, situational specificity contribues relatively little to the assessment of abilities.

Item Response Theory and Adaptive Testing. In discussing technical developments in test construction methodology, I chose validation as my first major example because of its pervasive implications for almost any practical use of tests. Although now undergoing considerable expansion, the concept of test validation itself has been familiar for a long time. My second example, in contrast, is probably the most conspicuous feature of the 1980s themselves and one that is likely to have a far-reaching impact on testing in the future. I refer to item response theory and adaptive testing.

Statistical techniques of item analysis have played an important part in test construction since the early days of standardized testing. These techniques have traditionally been concerned with the measurement of difficulty level and discriminative value of items. The first is based on percentage of persons giving the correct response, which is usually converted to a sigma distance from the normal curve mean. The second is based on the difference in total test score (or score on some external criterion) between those passing and those failing the item; this relation is often expressed as a biserial correlation.

It is apparent that both types of item measures are restricted to the samples from which they were derived and are generalizable only to populations that these samples adequately represent. For many testing purposes that require sample-free item measures, the procedure employed until recently was some variant of Thurstone's absolute scaling. This procedure requires the inclusion of a set of common anchor items across any two samples, in order to work out a conversion formula for translating all item values from one sample to another. A chain of linked sample values can be employed whereby all item values are expressed in terms of one fixed reference group.

With the increasing availability of high-speed computers, more precise mathematical procedures are gradually being adopted to provide sample-free measurement scales. These procedures were originally identified under the general title of "latent trait models." There is no implication, however, that such latent traits exist in any physical or physiological sense, nor that they cause behavior. They are statistical constructs derived from empirically observed relations among test responses. A rough initial estimate of an examinee's latent trait is the total score obtained on the test. In order to avoid the false impression created by the term "latent trai ," some of the leading exponents of these procedures have substituted the term "item response theory" (IRT), which is now gaining usage within psychology.

By whatever name they may be called, these procedures utilize three parameters: item discrimination, item difficulty, and a lower-asymptote or "guessing" parameter corresponding to the probability of a correct response occurring by chance. Some simplified procedures, such as the Rasch model, use only one parameter, the difficulty level, on the assumption that item differences in the other two parameters are negligible or can be eliminated by discarding items. This assumption, however, has to be empirically verified for each test. IRT is gradually being incorporated in large-scale testing programs, such as the College Board's Scholastic Aptitude Test.

One of the most important applications of IRT is to be found in computer-administered adaptive testing (CAT), also described as individualized, tailored, and response-contingent testing. This procedure adjusts item coverage to the responses actually given by each examinee. As the individual responds to each item, the computer chooses he next item on the basis of the individual's response history up to that point. Essentially each person takes a test that is tailor-made to fit his or her performance. The test stops when enough information is available to reach a preestablished reliability; this is equivalent to reducing the error of measurement to an acceptable level. The person's test score is based, not on the number of items passed, but on the predetermined score of each of the items passed, as determined by its difficulty level discriminative value, and susceptibility to guessing. The item "score" represents the best estimate of the ability level at which the likelihood of passing the item is 50-50. Adaptive testing is thus made possible by the use of

IRT in developing the item pool.

Exploratory research on computerized adaptive testing has been in progress in various contexts. Its operational use is under investigation in several large-scale testing programs, of which the most impressive is the CAT-ASVAB, or computerized adaptive version of the Armed Services Vocational Aptitude Battery.

Substantive Interpretation of Test Scores

So far I have touched on two of the three major areas I planned to discuss, namely role of the test user and technical methodology of test construction. Let me now turn to the third, the substantive interpretation of test scores. In contrast to the statistical interpretation of test scores, their substantive interpretation requires knowledge about the behavior domain assessed by the test. Many current misuses of tests and misinterpretations of test scores result from the application of erroneous or outdated knowledge about human behavior. I shall take illustrations from the intellectual domain, where misconceptions about the meaning of test scores are especially prevalent.

Nature of Intelligence. Let us consider, for example, the nature and composition of intelligence as investigated by factor analysis. The controversy over Spearman's *g* versus the group factors or separate aptitudes proposed by Thurstone and others flourished in the 1920s and 1930s. Recently this controversy has been revived and has received considerable attention in the popular media, which thrive on controversy.

In trying to work our way through this tangle of conflicting claims, we should bear in mind at least two points. First, the general factor identified in any one battery has often been loosely described as Spearman's *g*, suggesting a comprehensive general ability that underlies all intellectual activity. Actually, it represents only the general factor common to the tests in that battery. To conclude from such an analysis that a given test is heavily loaded with Spearman's g is misleading. It would be more meaningful to say that the general factor identified in that battery is heavily loaded with what that test measures, and this can be specified by examining the content of that test — for example, verbal comprehension, mechanical aptitude, or whatever. This is what we normally do in naming any factor identified in a factor analysis — we look at the test or tests in which the factor is heavily loaded, and we name the factor accordingly. Why not follow the same practice in naming a factor common to the whole battery?

My second point pertains to why factor analysis is conducted. Factor analysis is no longer regarded as a means of searching for the primary, fixed, universal units of behavior, but rather as a method of organizing empirical data into useful categories through an analysis of behavioral uniformities. Like the test scores from which they are derived, factors are descriptive, not explanatory; they do not represent underlying causal entities. Once we recognize the descriptive nature of factors, we see that the description could occur at different levels. More and more, we are coming to think in terms of a hierarchical model of factors or abilities: at the top is a general factor; at the next level are broad group factors, similar to some of Thurstone's primary mental abilities; these major group factors subdivide into narrower group factors at one or more levels; the specific factors are at the bottom level.

Different theories focus on one or another level of this comprehensive hierarchical model. No one level, however, need be regarded as of primary importance; rather, each test constructor or test user should select the level most appropriate for her or his purpose. This solution is what is actually done in practice. For example, if we want to select applicants for a difficult and highly specialized mechanical job, we would probably test fairly narrow perceptual and spatial factors that closely match the job requirements. In selecting college students, on the other hand, a few broad factors such as verbal comprehension and numerical reasoning, would be most relevant.

Another approach to the analysis of intelligence is that followed by cognitive psychology. This is a recent and rapidly spreading development in psychology as a whole. From the standpoint of testing, the principal contribution of cognitive psychology is its concern with what the individual does when performing an intellectual task. Their research concentrates on the processes rather than the products of thinking. Test performance typically assesses the products, as reported in test scores. To be sure, interest in processes is not new in the history of psychometrics. But the cognitive psychologists have carried the techniques of process analysis to new heights of refinement and sophistication.

Information about the processes an individual uses in solving problems or performing intellectual tasks is especially useful in diagnostic testing, to pinpoint the source of an individual's difficulties. It is also highly relevant in designing training programs to fit individual needs. For personnel selection and classification purposes, however, tests constructed through the traditional correlational and factor-analytic methods are still proving the more effective -- and there are sound theoretical reasons for this finding. Although some psychologists have seen controversy and conflict between traditional and cognitive approaches, again we find a movement toward a morè comprehensive view that incorporates both approaches, each serving different practical needs.

Testing in Context. The examples I have cited demonstrate that current knowledge about the nature of intelligence is certainly relevant to the proper use and interpretation of test scores. I now want to introduce one more topic that I consider even more important, which I shall call testing in context.[2] Test scores tell us how individuals perform at the time of testing, not why they perform as they do. To find out why, we have to consider the test score within the person's antecedent context. We need to delve into the individual's reactional biography. In what environmental setting did this person develop? What conditions and events were encountered and how did the person respond to them?

From another angle, we need to consider the test score within the person's anticipated context. What is the setting -- educational, occupational, societal -- in which this person is expected to function and for which he or she is being evaluated? What can we find out about the intellectual, emotional, and physical demands of that context? Several concepts encountered in the recent psychological literature, such as functional literacy and the assessment of personal competence, arise from this approach to test interpretation (Sundberg, Snowden, & Reynolds, 1978). The full understanding and proper interpretation of a test score has both a backward and a forward reference to real-life contexts.

It is now widely recognized by psychometricians that all cognitive tests measure developed abilities, which reflect the individual's learning history.

[2] For a provocative theoretical analysis of the role of context in thinking and intelligence, see the discussion of "situational cognition" and " personal and social epistemologies" by Greeno (1989).

This is equally true of tests traditionally labeled aptitude tests and those labeled achievement tests. The two types of tests differ principally in the degree to which the requisite prior learning is specified and controlled.

If we think of tests as measuring developed abilities, we can reformulate questions about *test coaching* in more meaningful terms. The basic question is not how far test scores can be improved by special training, but how such improvement relates to intellectual behavior in real-life contexts, such as performance on a job or in a course of study. To answer this question, we must differentiate between coaching that is test specific and coaching that affects the broader area of performance that the test is designed to assess, that is, the criterion we are trying to predict (Anastasi, 1981; 1988, pp. 43-47). Any condition that alters test performance -- for better or for worse -- without correspondingly affecting criterion performance will simply lower the validity of the test and make it a poorer predictor for the individual concerned. However, when coaching improves *both* test performance and criterion performance, it leaves validity unchanged, while enhancing the individual's chances of attaining desired goals. This broadly oriented type of so-called coaching could be more appropriately described as a form of short-term, condensed education.

Reformulating the coaching question in terms of the relation between test performance and criterion performance is also helpful in examining the widely debated question of *test bias*. The goal is for tests to be free from cultural bias against any group with which the tests are used. This does not mean that there can be no group differences in test scores. Such differences could correctly reflect differences in antecedent development of the skills and knowledge covered by the test, which may also be required for the criterion performance that the test is designed to assess — in a course of study, a job, or other real-life context. Essentially, a test is free from bias and equally fair to two groups if it has the same validity for both groups and does not underpredict the performance of either group. In terms of the familiar regression model, this refers to the avoidance of slope bias and intercept bias of the regression lines.

Try to visualize a scatter diagram or bivariate graph in which the horizontal axis shows test scores and the vertical axis criterion performance, both expressed in the same units, such as standard scores. Under these conditions, the slope of the regression line is exactly equal to the correlation between test and criterion, that is, the validity coefficient. If we plot the results of a minority and a majority group on the same graph, the two regression lines should have the same slope when the validities are the same for both groups.

Even with equal validities, however, there could still be intercept bias, if the two lines intersect the criterion axis at different points. This could mean that a minority person getting a lower test score than a majority person might perform equally well on the criterion. In other words, the minority test scores would underpredict criterion performance. Empirical studies have actually found the reverse. It is generally the group that scores higher on the test that tends to be underpredicted. There is sound statistical reason for this finding: as more tests are added to the battery, each of which has some predictive validity, the underprediction disappears (Linn & Werts, 1971; Reilly, 1973). The underprediction in the higher scoring group is likely to occur if the two groups differ in one or more additonal variables that correlate positively with both test and criterion. In any event, whatever the empirical findings in particular instances, checking for slope and intercept bias is the appropriate procedure for assessing the presence or absence of test bias.

From a broader viewpoint, all testing should be considered within a framework of cultural diversity and evaluated within its appropriate context. No test is -- or should be -- culture-free, because human behavior is not culture-free. We live in a pluralistic society, not only within large, heterogeneous nations such as the United States, but also within the broader, world-wide society. Increasing international contacts require some reconceptualization of mental measurement. Each test should be fitted into this broad framework. For practical testing purposes, the most effective tests are likely to be those developed for clearly defined purposes and for use within specified contexts. Although these contexts will vary in breadth, none is likely to cover all testing purposes nor the entire human species. The important point is to identify the locus and range of cultural (or other experiential) context for which any given test is appropriate, and then to keep both the use of the test and the interpretation of its scores within those contextual boundaries. In other words, when selecting or developing tests and when interpreting scores, consider context. I shall stop right there, because those are the words, more than any others, that I want to leave with you: consider context.

References

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*, 1086-1093.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1-15.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anastasi, A. (in press). The test user qualification project: An evaluation. *American Psychologist*.

Balma, M. J. (1959). The concept of synthetic validity. *Personnel Psychology*, *12*, 395-396.

Bentz, V. J. (Chair). (1980, September). *Methodological implications of large scale validity studies of clerical occupations* Symposium presented at the annual convention of the American Psychological Association, Montreal.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primroff, E. S., & Most, R. B. (1988). Test user qualifications: A data-based approach to promoting good test use (Report of the Test User Qualifications Working Group of the Joint Committee on Testing Practices). *Issues in Scientific Psychology*. Washington, DC: American Psychological Association.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B. (in press). Test user qualifications: A data-based approach to promoting good test use. *American Psychologist*.

Greene, J. G. (1989). A perspective on thinking. *American Psychologist*, *44*, 134-141.

Kleinke, J. D. (Ed.). (1987). Edison Electric Institute employee selection testing project: Consortia that work. Washington DC: Edison Electric Institute.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, *8*, 1-4.

Reilly, R. R. (1973). A note on minority group test bias studies. *Psychological Bulletin*, *80*, 130-132.

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., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*, 473-485.

Sundberg, N. D., Snowden, L. R., & Reynolds, W. M. (1978). Toward assessment of personal competence and incompetence in life situations. *Annual Review of Psychology*, *29*, 179-221.

Turban, D. B., Sanders, P. A., Francis, D. J., & Osborn, H. C. (1989). Construct equivalence as an approach to replacing validated cognitive ability selection tests. *Journal of Applied Psychology*, *74*, 62-71.

MANAGEMENT ISSUES IN THE NEXT DECADE: IMPLICATIONS FOR THE HUMAN RESOURCE MANAGER

NEWTON MARGULIES, PH.D.
UNIVERSITY OF CALIFORNIA, IRVINE

AND

ANTHONY P. RAIA, PH.D.
UNIVERSITY OF CALIFORNIA, LOS ANGELES

INTRODUCTION

Initially, this research project was focussed on generating information about the current status of the American corporation. We were interested in investigating the ways in which organizations postured themselves toward global changes, including economic changes, social changes, and cultural changes. One element of our literature review, for example, had indicated that the core values of the field of organizational development seemed to have remained relatively constant over the past twenty years. And yet some analysts and theoreticians speculated about the conservative and even anti-humanism that seems to be the trend in organizations today. Our interest, therefore, was to examine the compatibility of organizational development as a field, and the directions of American organizations.

Our data gathering began with this focus and continues in this vein as our interest in the values of American organizations grows. Additional insights, however, into the critical issues facing organizations in the near future and what organizations are currently doing to address these issues provides some important information for policy and activity formulation.

RESEARCH DESIGN

The basic design of this research utilizes a combination of questionnaire and in depth interviews with managers in a variety of organizational settings. Phase I of this project began with an interview phase in which we explored with managers the issues they perceive will confront their organizations and them as individuals in the coming decade. We also asked, as part of our exploration, what specifically their organizations were doing or planning to do in preparation for addressing and resolving the issues they identify. Finally, since our initial research began with an exploration of corporate values, we continued to explore the arena of values and ethics through our interview procedure, and specifically, through the use of a relatively short semantic differential questionnaire.

As we proceeded with our interviews we were able then to design an open-ended questionnaire in which managers could write responses to such questions as: Identify the issues confronting your organization in the coming decade? What specifically, is your organization doing to address these issues? These open-ended questions, six in total, were then content analyzed to create categories so that responses could be easily catalogued.

We continued our follow-on interviews with managers, exploring in depth the nature of the issues they identified and the potential impact these issues might have on organizations as well as individuals.

To date, our sample includes 456 managers in industries which include aerospace, biotechnology, retailing, health care, electronics, software development, manufacturing, and some public agencies.

Additionally, this particular paper compares the results of our study to date and the results of a study conducted by the Conference Research Board in which human resource managers identified the issues facing their function in the coming years.

RESULTS

The analysis of our data indicates the following issues, as identified by managers. For the most part a high percentage of the respondents indicated that fierce competition spurred by new perspectives on the global economy was a major issue facing their companies. In addition, issues such as the quality of products and services, the speed of product innovation, responding to market differentiation as well as new market development, and continued concerns for productivity and efficiency top the list.

On the organizational side, a high percentage of managers indicate that issues surrounding the restructuring of their organizations, and perhaps more importantly, the implications for major culture change will be highlighted in the coming decade.

Other issues reflecting changes in the relationship with government agencies, changes spurred by continued interest in mergers and acquisitions, and changing management-employee relationships were articulated in many of our interviews.

To summarize, the data clearly indicates that major issues facing organizations in the coming decade evolve around a growing intensity and need for change. This includes not only a change in thinking about economies, products, and market place, but also serious concerns about the changing organization and its relationship with its environment, including government, legal, and culture.

In response to the question: What is your organization currently doing to respond to project issues for the next decade?, managers indicated the following activities:

Toping the list are activities like: learning to do better strategic planning, developing a new management philosophy and style, continued training in both technical and managerial arenas, more careful recruiting considering future needs, learning to manage teams and interfaces more effectively, and in many cases, using consultants in very specific and specialized areas.

Additionally, managers also indicate that serious thinking about reorganizing and streamlining their organizations, as well as learning to manage teams and interfaces more effectively are important considerations when thinking about the next decade. Most importantly, there are concerns about the quality of leadership and the nature of that leadership required in the next decade. The identification of new strategies for coping with bureaucracy is also a major concern.

Clearly, our data indicates that managers recognize the need for new skills in the leadership and management arena, and continued employee development via training in both the technical and supervisory areas. Most managers are quite aware that being cognizant of new technology and rapid technological development must be on the forefront of their role in the future.

We also asked whether or not there were other things their organizations are not doing but should be doing. Responses indicate that there is a need for longer range planning, better management of change, quicker turn around on product development, and much better coordination between groups and functional departments.

On the more personal level managers see their own jobs as being more complex and more stressful in the coming decade. They project an increase in work load and continued concern about the obsolescence of their skills.

IMPLICATIONS

In examining this information, the implications for human resources seems obvious. For example, managers are concerned about the need for skill development, as well as assistance in the management of change. Both the line manager population, as well as HR managers concede that focus on organizational culture change may become a central issue of the next decade. Additionally, it would seem that line managers are vitally concerned about ongoing skill development, in both technical and managerial areas. While human resource managers indicate that training is of concern, the priority put on this issue by these groups is quite different. The line managers rank training much higher than do their HR manager counterparts.

Many line managers indicate that there is indeed a need to review and manage the changing relationship between management and employees. Many organizations reflect a deepening cynicism which currently characterizes the relationship. HR managers, however, view involvement and participation considerably lower on the priority list than do line managers. Both groups agree that the management of teams and interfaces will become more important in the next decade and that there is a need to become more skillful and adept in this arena.

We believe the implications of our study in contrast to the data provided by the Conference Research Board indicate the need for HR to better understand the concerns of the line manager. The data from HR managers indicates that the function is not as close to the core business as it would believe, and does not have impact on the success of the organization that it should.

THE GUIDELINES IN THE YEAR 2000

Lance W. Seberhagen, Ph.D.
Seberhagen & Associates*

In October 1985, I was invited to testify before the House Subcommittee on Employment Opportunities regarding on what changes, if any, should be made in the Uniform Guidelines on Employee Selection Procedures (EEOC, 1978). This paper is based on that presentation, but a lot could happen between now and the year 2000 to change my recommendations. Of immediate concern is the Supreme Court's decision in Wards Cove, but there will undoubtedly be other important new developments in psychology and the law that must be woven into any general standards or guidelines for employee selection.

Should the Guidelines Be Revised?

The net effect of the Guidelines has been a very positive one for the development and use of good employee selection procedures and the reduction of employment discrimination. Therefore, any change in the Guidelines s:ould be viewed with caution. If important aspects of the Guidelines are weakened or eliminated, I would be opposed. If important aspects are retained in the Guidelines, I would like to see further revision to promote equal employment opportunity and improve workforce productivity through better employee selection procedures.

What Important Aspects of the Guidelines Should Be Retained?

1. Definition of "selection procedure" as "any measure, combination of measures, or procedure used as the basis for any employment decision."

2. Reliance on generally accepted professional standards (e.g., APA Standards, SIOP Principles) as the ultimate technical basis for employee selection.

3. Rejection of casual evidence of validity (e.g., qualifications of test developer, frequency of usage, testimonials).

4. Documentation requirements (e.g., technical report of validity, manual for administration and use) for each selection procedure and the overall selection process.

What Changes Should Be Made in the Guidelines?

1. Provide one set of guidelines to provide truly uniform federal policy on the development and use of employee selection procedures. This would require extending coverage of the Guidelines to age, handicap, veteran's status, and other bases of employment discrimination prohibited under federal law.

* 9021 Trailridge Court, Vienna, VA 22182. Tel. (703) 790-0796.

2. Provide guidance for all forms of discrimination (e.g., unequal treatment), not just adverse impact.

3. If the adverse impact concept is retained, re-define "adverse impact" to require a combination of the 80% Rule and statistical significance.

4. Require all selection procedures to be valid, regardless of adverse impact.

5. Provide more guidance for proper job analysis. Any method of job analysis should be acceptable if it is valid, reliable, fair, and useful for the intended purpose.

6. Provide more guidance for validity generalization based on "test transportability."

7. Provide more guidance for validity generalization based on the meta-analysis of validation studies.

8. Require analysis of the total selection process to ensure that all necessary worker characteristics are assessed in a proper way to ensure job success.

ARITHMETIC VERSUS CLINICAL PROCESSES FOR DERIVING ASSESSMENT CENTER SCORES

Patrick T. Maher, Principal Associate
Personnel and Organization Development Consultants, Inc.
5842 Crocus Circle, La Palma, CA 90623 (714) 827-1780

A continuing controversy, fueled by contradictory findings, centers on whether or not assessment center scores should be derived arithmetically or through clinical judgement.

Sackett and Wilson (1982) developed a decision rule that would predict final consensus scores with 94.5 percent accuracy.

Joiner and Carlin (1983) reported to IPMAAC that a comparison of pre-integration dimensions scores and post-integration scores produced a correlation of .98 and identical rank order lists in a selection assessment center for law enforcement. They replicated their work in 1985 and reported to IPMAAC that high correlations (.92 or better) were again achieved comparing pre-integration and post-integration scores, but these demonstrated changes in rank-order positions that in some cases were significant.

Lowry (1988) reported that similar studies showed differences in post- and pre-integration rank order lists for career development assessment centers, but not for selection assessment centers.

In summary, although the research shows mixed results, it seems to indicate that arithmetic scoring, at least in employment selection assessment centers, is a viable alternative to the lengthy discussion involved in the clinical process. (E.g., a consensus discussion time of one to three hours per candidate are somewhat typical.)

Keep in mind that the prior work, however, had been largely restricted to comparing pre- and post-integration and rank-orders based assessment center results alone. The research presented here analyzes data obtained when the assessment center or assessment center simulation is a weighted part of a multi-phased assessment procedure.

It should be noted that there are a variety of assessment center rating systems in use. Thus, a brief description of the various processes used in these particular assessment centers is important presenting results, data, and findings.

All dimensions were rated by three assessors assigning a score of one (low) to five (high) in each. Where there was only one exercise, overall dimension scores were used.

However, where there were several exercises, overall dimension scores were derived slightly differently. After assigning individual scores by exercise within dimensions, overall scores were then assigned for each dimension. However, these overall scores were not a mere averaging of exercise scores, but required the assessors to consider a variety of factors such as the effectiveness of one exercise compared to another in measuring a dimension (E.g., a written simulation requiring lengthy narrative is usually considred more valuable for measuring written communication than the in-basket. However, in some cases, a particular in-basket might emphasize writing skills or a particular candidate may provide a better example of written communication in the in-basket; therefcre that exercise will be given greater weight in this dimension.)

In assigning the scores, each of the assessors initially made independent ratings of the scores. When they met for assessor discussion, each assessor stated his or her score orally, and any difference in scores was discussed, even if it was only a one-point difference. It should be noted that this varies from a common practice of holding a discussion only if there is a difference of 2 or more points in assessor scores. (This two-point difference was used by Lowery, which may account for the difference between his findings and the findings here).

The dimension scores were then arithmetically computed to determine an overall score on a scale of 50 to 100. This mathmatic computation merely converts the clinically-derived scores into a scale of 100, as required by most civil service or merit systems. This computation should not be confused with other systems designed to arithmetically derive scores without assessor discussion.

The results of two different examination procedures used for two different positions are presented in this research.

The first examination was for the position of purchasing agent. In most jurisdictions, a purchasing agent's limited to preparing requests for proposals, analyzing projected costs, and, in some cases, making recommendations for selecting vendors.

In this particular jurisdiction, however, the purchasing agent acted more as a manager-executive. As such, he was required to perform a number of administrative and executive functions, and could serve as an interim department head.

Therefore, quite a few of the candidates applying for the position did not have the requisite supervisory and management experience, regardless of their length of service as a purchasing agent in other jurisdictions. As a result, we noted that there were a large number of low scores derived from assessment procedures. This was not inconsistent with the experience level of the candidate pool in this particular assessment procedure.

The assessment procedure consisted of two components, each weighted at 50%. One component consisted of the General Management In-Basket (GMIB), developed and marketed by Management Personnel Systems. This in-basket is a standardized assessment procedure, assessing leadership style and practices, priorities/sensitive issue handling, conflict/interpersonal insight management, and organizational practices/management control. These four factors result in a total score used to determine the candidate's ranking.

The other half of the assessment procedure consisted of a written simulation in which twenty pages of statistical and narrative data were presented to the candidates. Candidates were then required to analyze the data and develop recommendations. They prepared a narrative staff report incorporating their summary and data analysis, and their recommended courses of action. This portion of the assessment procedure evaluated written communication, problem analysis, and decision-making.

A second part of the simulation exercise directing the candidates to make an oral presentation of their report. This allowed the assessment of oral communication skills in what was essentially a written simulation exercise.

Strictly speaking, this process did not constitute a true assessment center because of the limited number (2) of simulation exercises. However, all other aspects of the assessment center process were followed, including the use of raters. (These were trained for three days and required to demonstrate assessor competency on the fourth day.)

The second examination pertaining to th research was , for the position of battalion chief in a large metropolitan fire department. The examination process consisted of four parts.

A paper-and-pencil test, weighted at 15%, posed 100 multiple choice questions on technical fire-fighting knowledge and departmental procedures. A fire-scene simulation, weighted at 25%, involved an oral presentation and interaction in combating a simulated fire. An assessment center, weighted at 35%, assessed eight generic management abilities. It was comprised of an in-basket, group discussion, and written simulation exercise. A departmental evaluation, weighted at 25%, was based upon on the job performance. It was rated by battalion chiefs and deputy chiefs in the department.

After compiling all four examination components and computing the respective scores, seniority points were addeo to the overall score based upon the number of years of service time as a fire-fighter. This resulted in an overall score, which in one case exceeded 100 on a scale of 100.

The results of the two examinations are presented in the attached tables. Table 1, shows the rank-order data for the purchasing agent examination. The pre- and post-simulation scores show the scores for simulation exercise. The pre-score represents the pre-consensus discussion and the post-score the post-consensus discussion. The overall scores consist of the simulation score combined with the GMIB score. The pre- and post-overall scores represent the GMIB combined with the pre-consensus and post-consensus scores.

Table 2 shows the rank-order data for the battalion chief examination based on the pre-post-consensus scores for the assessment center. The overall scores (all components) are also included, but additional data are presented. The "NACOvr" score is the total score for all components except the assessment center. Itdemonstrates the rank order without any of the assessment center scores.

The assessment center score is the actual assessment center score based on the overall dimension score. The pre- and post-consensus scores are based on the individual exercise scores.

As can be readily seen by comparing the PstCon score with the assessment score rank order, the two are not identical, although a correlation of .98 is obtained.

The difference in the rank order, based on the identical scores but obtained through different scoring procedures, tends to support Lowry's contention (1988) that the scoring method used affects the assessment center results.

As can be seen, there is no difference in the scores of the top three rank-ordered candidates in either examination, or those of the top four rank-ordered candidates for the purchasing agent examination. These results may be reflective of Lowry's finding with his small candidate populations (n=4 on selection and n=6 on career development).

However, when the larger numbers are examined , we see a change in the rank order of candidates, in some cases, a significant one. More importantly, rank order changes when it involves the overall score.

Sackett and Wilson's work is also impacted in light of the rank-order results. While they can predict with 94 percent accuracy the post-consensus discussion results, rank order is affected. And, in the public arena, where strict rank order often determines not only when a candidate is promoted, but if he/she is promoted, such accuracy has a tremendous personal impact.

As indicated above, all of the research, including the current studies, reports high correlations (.92 - .98) between scores obtained through independent assessor ratings and consensus ratings. However, such correlations are not relevant to the issue of whether or not scores should be obtained through arithmetic or consensus discussion.

The issue, especially where promotion is dependent entirely or primarily on rank order of the list, is the extent to which rank order is affected when mathematical comuptation or consensus discussion is used. While rank order may not always be affected, as demonstrated by Lowery and by Joiner and Carlin, nonetheless, rank order can be affected. This change is not only demonstrated by the research presented here, but by Carlin and Joiner in their replication study and Lowery in his-career development assessment center.~

One disturbing contention by Lowery and Sackett and Peters is that consensus discussion is more valuable in career-development assessment centers than in selection assessment centers, at least as far as its impact on rank order. In reality, rank order in career development assessment centers is not as important as in selection assessment centers. Rank order in career development assessment centers is beneficial only in demonstrating standing relative to others. Where rank order is based on close scores (e.g., 93.5 and 93.0), it has no value.

In promotional assessment centers, however, rank order can mean the difference between whether or not a candidate is promoted, even if the difference is only a hundreth of a point. Besides determining if a candidate will be promoted, it determines when the candidate will be promoted. In many cases, an early promotion not only means additional benefits amounting to thousands of dollars, but could affect future promotions at even higher levels.

In summary, it seems that the research on change or lack of change in rank order has not really addressed the critical issue. That is, whether consensus discussion results in a more valid and reliable score than does the arithmetic process. Future research, especially criterion research, must explore this aspect of the issue..

TABLE I
RANK ORDER OF CANDIDATES
BY PRE-CONSENSUS/POST-CONSENSUS RESULTS
AND BY GMIB

Post-Over	Pre-Over	Post-Sim	Pre-Sim	GMIB
1	1	1	1	2
2	2	4	4	1
3	3	3	3	4
4	4	2	2	12
5	6	8	10	3
6	5	5	5	5
7	7	7	6	6
8	8	6	7	8
9	11	9	9	10
10	10	11	11	7
11	9	10	8	11
12	12	13	12	9
13	13	12	13	13

Correlations:	Pre-sim	Pre-ovr
Post-sim	.951	
Post-ovr		.980

TABLE II
RANK ORDER OF CANDIDATES
BATTALION CHIEF EXAMINATION
BY PRE-CONSENSUS/POST-CONSENSUS RESULTS

Over*	AsmtCn*	PreCon	PstCon	PreOvr	PstOvr	NACOvr*
1	1	1	1	1	1	1
2	2	3	2	2	2	2
3	3	2	3	3	3	4
4	7	5	6	4	4	3
5	8	11	11	5	5	5
6	6	7	10	6	8	6
7	4	4	4	7	6	9
8	5	6	5	8	7	8
9	9	9	9	9	9	7
10	11	8	7	10	10	12
11	13	13	13	12	12	10
12	12	12	12	11	11	11
13	10	10	8	13	13	13

Correlations

	PostCon	PostOvr
PreCon	.977	
PreOvr		.993

* Actual test scores. "Overall" is the actual ranking of the candidates involved in the examination process.

REFERENCES

Joiner, D.A. and Carlin, P. "Consultant-Agency Cooperation in Conducting Research on a Promotional Assessment Center for Police Lieutenant." In C. J. Lindley, Publications Director, **Proceedings of the 1983 IPMA Assessment Council Conference on Public Personnel Assessment.** Washington, D.C., May 22-26, (1983), 39-40.

Joiner, D.A. and Carlin, P. "Replicating Research on Police Promotional Assessment Centers." in C. J. Lindley, Publications Director, **Proceedings of the 1983 IPMA Assessment Council Conference on Public Personnel Assessment.** New Orleans, LA, June 16-20, (1985) 39-40.

Lowry, P. E., "The Assessment Center: Pooling Scores or Arithmetic Dcision Rule?" **Public Personnel Management**, Vol 17, No. 1, (1988), 63-71.

Sackett, P. R., and Wilson, M. A., "Factors Affecting the Consensus Judgment Process in Managerial Assessment Centers." **Journal of Applied Psychology**, 67, No. 1 (1982), 10-17.

PRACTICAL APPLICATIONS TO SAVE TIME, MONEY AND HEADACHES IN JOB SIMULATIONS

by Jeff Prewitt, Chief Examiner, Louisville Civil Service

This paper provides tips to save money on assessor travel, lodging and the assessing site. Time saving procedures are discussed for everything from job analysis to assessor training. Headaches such as scheduling dilemmas, rumors, and complaining assessors are also addressed. Lastly, suggestions are made for inexpensive perks for assessors.

Saving Money

"We don't have enough money! Job simulations are too expensive." Job simulations can be expensive; however, some of the costs can be reduced with planning, investigating and negotiating on the three most expensive areas: assessor travel, assessor lodging, and the assessing site.

Most airlines give discounts if you book flights far enough in advance and flights are considerably less expensive if the traveler is staying over on a Saturday night. You can also save money on air fare by shopping around and finding which cities have cheaper air fares to your city. Another way to cut costs on assessor travel is to find agencies within driving range of your city and reimburse the assessors for their mileage. Sometimes you can look even closer than cities within driving range-try local agencies and local universities. These sources will not only save you money on assessor travel; they may save you money on assessor lodging as well since local people will not need hotel lodging.

Using local sources for assessors is a good way to reduce your costs for assessor lodging, but don't stop there. Ask hotels about government discounts and tax exemptions. Check on off-season rates. Shop around for the most economical lodging and ask for group discounts. If you are going to use a hotel for the assessing site, check to see if the hotel will give you discounts on the assessing rooms since you are using a large group of rooms and they will not need linens and towels. Also, ask if extra conference rooms, administrative rooms, or other perks are free when a large number of guest rooms are rented. You can even negotiate on catering lunch to save money, give you one less detail to attend to and to insure that none of your assessors get lost during lunch. Once you have squeezed everything you can out of the hotel to save money on assessor lodging and the assessing site, you can still save money on lodging by having your own staff conduct feedback sessions and using local assessors for the more time-consuming exercises. That way you can get your out-of-town assessors out of town sooner!

If you still have not saved enough money, try to avoid using a hotel as an assessing site. Look for training academies, universities, schools, agency buildings, and conference rooms or offices not being used. If you plan far in advance, you may be able to reserve several rooms. Another way to get more office space for assessing the candidates is to conduct some of your testing on the weekend when more office space is available. If equipment is a problem, contact other departments and see if they will loan video equipment to you and check with local universities to see if they have equipment and recording studios that you could use for a small price.

Saving Time

Planning in advance is the best way to save time. Develop a comprehensive checklist that covers all details, no matter how minor. Use this checklist for delegating, monitoring, and identifying things to take care of when you find yourself at a standstill. The first section of your checklist should cover the job analysis. When conducting your job analysis, contact other jurisdictions and request task and KSA lists from them. These lists can be used to: supplement your lists; help insure thoroughness; and check for acquiescence by Subject Matter Experts (SMEs). Additionally, critical incidents forms can give you a great starting point for job simulation scenarios and they may help you identify important tasks or knowledges, skills, and abilities (KSAs) which were missed with the other procedures. Lastly, if you find that your SMEs are having trouble dimensionalizing statements and responses, have your staff members dimensionalize and then work with your SMEs to insure accuracy.

Once your job analysis is complete and you have a test plan, divide tasks up on the checklist so that people are taking care of administrative details as well as exercise development details. Plan your schedules out in detail so that you are using all of your resources to maximum potential.

In addition to saving time in developing and administering the job simulation exercises there are steps you can take to save on the time used by assessors. Mail information packets to the assessors in advance which explain some of the procedures and the dimensions being measured. Include information on dining, entertainment, schedules, and the type of dress required of the assessors. Make your first impression one of professionalism, preparedness, and hospitality. After your assessors have arrived and gone through general training as a group, divide your assessors into groups so that the assessors only need to be trained on how to evaluate one type of exercise. Eliminating the consensus approach is another way to reduce assessing time.

We have also found that the scales are easier for the assessors to use when staff members serve as additional assessors. The staff members are also able to give excellent feedback at a later date so the outside assessors only focus on recording and evaluating behavior instead of the additional duty of giving feedback.

Reducing Headaches

After using the above mentioned pointers to reduce time and money spent on job simulations, you can direct your attention to some methods of reducing headaches caused by scheduling problems, rumors and complaining assessors. The key here is to anticipate problems and address them before they occur.

Most scheduling problems can be eliminated by a well thought out schedule that incorporates catch-up times at different intervals in the day. To lessen the impact of late candidates, schedule them to appear earlier than they will actually begin their exercises. If you administer the job simulation and videotape the candidates prior to the arrival of your assessors, you will not have to face the grueling task of coordinating candidates and assessors and rushing assessors so they are ready for the next candidate who is waiting outside their door.

Videotaping candidates in advance not only reduces headaches caused by scheduling dilemmas, it also allows the assessors to work at their own pace which is one way of reducing complaints from assessors. You must be sure to give your assessors enough time to evaluate candidates fairly and you should ensure that you do not overwork the assessors. Having an alternate assessor to give breaks to your assessors and to be available in case you have a problem with an assessor is also a good idea. Some other ways to avoid complaining is to ask the assessors on the first day if they had any problems with their rooms so you can address the problems early and take steps to ensure the assessors' comfort. You should also provide tourist and dining suggestions to the assessors and attempt to provide some entertainment or any types of perks you can arrange.

Perks for assessors can come in many forms, unfortunately you will probably have to look for inexpensive perks. The first place to look is the agency you are serving. Ask for badges, patches, hats, t-shirts, or lapel pins. You can also set up a hospitality room with support from the FOP or other local employee organizations. Sometimes, you may find local businesses who will sponsor activities or provide free tickets or discounts. Finally, certificates of recognition and "thank you" letters to the assessors and their superiors are nice touches to show your appreciation. They also help out the next time you ask for assistance from that agency.

The last group of headaches to address are what I call the GRIM Reapers (Grapevine, Rumors, Ignorance, and Myths) because they can kill all your other efforts. If you can dispel the rumors & myths and educate the candidates on the job simulation process, you can prevent a lot of problems. The three keys here are to inundate the candidates with accurate information, make your testing staff highly visible, and ask everyone for suggestions for improvement.

Some ways to increase the visibility of the testing staff are: conduct observation interviews during all shifts and in all the different divisions of the department; attend special training programs; and carry a beeper so you are notified of major fires. Your credibility increases and you learn more when you show up at major incidents to observe and ask questions. Lastly, attend question and answer sessions to make yourself more visible and to give out accurate information.

In addition to question and answer sessions, you should consider using videotaped descriptions of the job simulation process with "practical pointers" for disseminating information. The candidates will watch the tape for the practical pointers and you can sneak in some good education on the job simulation process from conducting the job analysis to the determination of final ratings. You can also use feedback sessions to educate candidates on the process and to solicit their suggestions on ways to make the process more job related and fair. This helps give you more credibility and you will actually get some good suggestions.

Benefits of Video Recording

Many of the suggestions provided in this paper require the use of video equipment. If you do not have video equipment available, ask for it. The benefits listed below should help you to justify video equipment at budget time.

-Removes intimidation by assessors.
-Insures consistent behavior by assessors.
-Allows exercises to be conducted prior to the arrival of the assessors cutting down on testing space and staff needed.
-Allows assessors to work at their own pace.
-Allows assessors to look at candidates' performance again if necessary.
-Provides an exact record of candidates' performance.
-Provides more detailed feedback for candidates.

Several procedures have been described which can be implemented to decrease the amount of time, money and headaches involved with the job simulations. Other jurisdictions should find these approaches easy to implement and justify.

MAKING APPRAISALS EFFECTIVE BY LIMITING AVERSIVE ELEMENTS

Daniel F. Twomey
Fairleigh Dickinson University

The need for consistency and formalization in government agencies is evidenced in the their appraisal systems. A variety of appraisal systems are used and more often than not an agency's appraisal system is a complex combination of two or more of the basic systems. For example the VA uses a trait system combined with a substantial MBO system. Also there are different systems and variations within agencies. There are of course good reasons for each system and its variations. Common to most of the systems is a complex set of ratings and rankings that inevitably lead to making appraisal aversive to many of the appraisers and employees.

In an effort to make the assessments better, the negative consequences of those more articulated systems have frequently been overlooked or simply accepted as a necessary evil--something that was unavoidable. This paper addresses the aversive side of appraisal. The sources and dynamics that make appraisal unpleasant for the appraiser, demotivating to many employees, and ineffective in promoting organizational goals will be explored. A case will be made that by eliminating much of the sting from appraisal, better evaluations and increased motivation and acceptance of the system are achieved.

The struggle to improve performance appraisal systems has been ongoing and mostly unsuccessful in business and in government. Big businesses have followed a similar pattern as government with increasing complex systems and organization-wide commitment to formal assessment of employees at all levels. With evidence that the established systems were not accepted or viewed as effective by either supervisors or employees, some have made radical changes that for them have met with success. The common elements of these new approaches will be analyzed and applied to generic types of organizations.

PRINCIPLE - ARTICULATE OBJECTIVES OF PERFORMANCE APPRAISAL AND MATCH THE PROCESS WITH THE OBJECTIVE.

The performance appraisal process must be appropriate for the objective of the performanance appraisal. For example, the process for evaluation is inconsistent with the objective of development. Development requires substantial open information exchange to develop the understanding and commitment to change. The evaluation process of the supervisors assigning numerical ratings to various traits, behaviors, or outputs puts the employee in the role of defending past performance rather than building an understanding and commitment to future improvement. If the evaluation process is used for development it is using force to create insight. The pressure will prompt defensiveness and limit openness. The result is unnecessary and dysfunctional coerciveness.

Conversely if an open problem solving "development" process is used for establishing ratings (evaluation) there is confusion about the criteria, information, and who actually decides on the ratings. Also the participation may create the impression that the employee has a major role in deciding the evaluation. The result is the employee expects a mutual decision and feels used if he or she doesn't have a say in the outcome. If the employee is a high performer it may not be a problem, but for others, whose self impressions are more positive than the appraiser's ratings, it is alienating.

PRINCIPLE - LIMIT RATING AND RANKING TO THOSE THAT ARE CLEARLY NEEDED.

All appraisals are to some degree subjective and therefore are potentially coercive. In most systems there is either an implicit or explicit requirement for some average or below-average rating or ranking. If they are perceived as unfair or unnecessary they are coercive. The purpose of each and every rating and ranking should be carefully reviewed and only if it is absolutely necessary should the rating be continued. If merit pay rankings are needed, only the necessary ranking should be made, e.g. if there are three merit levels, persons should be placed in each level, but no further rankings should be made. If there is no merit system, routine rankings should be avoided. Overall or summary ratings can generally be omitted, since they don't contribute substantially to the informational feedback. Furthermore, the psychological impact of telling someone that overall he or she is unsatifactory can be devastating, and there is no precise behavior that can be identified, that should be changed, as the means of correcting the summary deficiency.

PRINCIPLE - DON'T USE INDIVIDUALIZED CRITERIA FOR RANKING, IF MERIT IS AWARDED COMPETITIVELY.

When a MBO system is used as part of the appraisal system it should not drive the appraisal unless it is fully funded in the sense that all meritorious performance gets rewarded based solely on the achievement of goals. In those situations where budgets drive the merit system, especially if it limits the number of possible recipients, the potential for dissatisfaction is high, and the basic tenets of MBO are being violated. In those situations some shared criteria would be better than individual goals in awarding merit. Individual goals may still be set and achievements measured, but since the system doesn't have the capacity for a direct linkage with goal achievement, pretending they do or forcing distortion into the system creates resentment.

PRINCIPLE - DON'T USE SELF RATING OR RANKING UNLESS THERE IS AN OBJECTIVE BASIS FOR THE APPRAISAL.

In some programs the employee is required or allowed to rate him/herself prior to the performance interview. The supervisor also rates the person. These ratings are then shared in the performance appraisal interview. In most cases substantial differences are likely to occur. The evaluation provides neither the climate nor the information to resolve the conflict in an integrative manner. The greater the complexity of the job and the uncertainty of the measurements the more difficult is any satisfactory resolution. The conflict is resolved either by forcing or by smoothing over. In either case the employee is likely to have been put down. With jobs that are routine and with job performance measurements that are factual, the dual rating may help identify incongruities between the supervisor's and the employee's perceptions which, due to the objective basis, may be effectively resolved. In a developmental performance appraisal program self appraisal is one of the most powerful means of creating a data base and a problem solving climate, but even in this setting actual ratings, e.g. against a numerical scale, are not recommended. A descriptive "rating" is better.

PRINCIPLE - DON'T LET THE PERFORMANCE APPRAISAL SYSTEM DEFINE A LARGE PERCENTAGE OF EMPLOYEES AS LOSERS.

Administrative decisions frequently require the selection of one or a few out of many. This and other administrative processes may influence the appraisal systems in ways that hinder the prime purpose of appraisal - to increase productivity. Motivation and productivity flow from positive reinforcements, positive self image, high expectations, and a supportive climate. Simply, empl.yees who feel like "winners" are better than those that feel like "losers". Many appraisal systems are designed to limit the winners to less than 50%, if winner is defined as being above average. The term average is one of the most ambiguous words and despite its formal definitions carries a negative connotation. Deficient performance needs to be identified, and outstanding performance rewarded, but the system should allow for all good employees to be so rated. Gellerman and Hodgson (1988) report on Cyanamid's reduction of forced low ratings is a good example of making appraisals more effective by reducing low ratings.

PRINCIPLE - USE REWARDS TO INCREASE MOTIVATION NOT TO CREATE UNDERSTANDING.

Evaluations and the rewards that flow from positive performance and evaluation, increase the motivation of the employees, but do little to increase the employees understanding of the job. Employees whose performance is blocked by lack of understanding of complex aspects of their jobs need additional self insight that comes from development. Increasing pressure to perform without creating a better understanding of how to improve performance increases frustration and does little for productivity.

PRINCIPLE - THE EMPLOYEE SHOULD UNDERSTAND HIS/HER RIGHT TO QUESTION AND THE SUPERVISOR'S RESPONSIBILITY TO JUDGE.

In many appraisal situations neither the appraiser nor the appraisee understands their role. When an employee believes he or she is being evaulated unfairly there is uncertainty of whether it is appropriate to argue one's case. The issue of whether the supervisor has a right to judge and the issue of whether the employee has the right to challenge the role and ratings of the appraiser is unclear. If the employee doesn't argue, he or she may feel disappointed in oneself, and if one does argue it may hurt the relationship with the supervisor. It should be established and understood that the supervisor has the responsibility to make reasoned judgements, and the employee has the right to know the process, criteria, sources of information, and how the ratings were derived. By understanding each other's role, useless win/lose arguements can be avoided.

PRINCIPLE - THE LACK OF MAJOR GRADATIONS IS COERCIVE TO HIGH PERFORMERS, AND FOR THE SUPERVISOR.

A common occurrence is that pressures within the system, such as having to justify poor ratings, or inadequate information or training for the supervisor, causes ratings to be high and not reflect the gradations in performance. This, of course, hurts the high performers by denying their relative contribution and the rewards that should follow. The supervisor who is caught up in such a system is likely to be punished if he or she substantially violates the norm of inflated rating.

Performance appraisals have several stakeholders each of whom must be satisfied if the system is to work. The employer needs a system that is integrated with other systems, that contributes to productivity, and is legal. Management needs a system that is not overly burdensome and one that supports managers and other management systems. The employees need a system that they believe is fair, leads to equitable rewards, and helps development and career advancement. It is the employee who is often overlooked. Many systems simply do not have the support of employees. Establishing an effective appraisal system is a most difficult task with no pat answers. There are many issues to be dealt with other than reducing the level of coerciveness. But if the unnecessary and avoidable coercive elements can be eliminated, the other issues will be easier to resolve.

REFERENCES

Gellerman, S.W. and W.G. Hodgson. "Cyanamid's New Take on Performance Appraisal", Harvard Business Review, 66(3), May-June, 1988, 36-41.

USING EVALUATION AS A TOOL FOR ENHANCING PROFESSIONAL DEVELOPMENT IN HIGHER EDUCATION: CONDUCTING THE FEEDBACK INTERVIEW

Ann F. Lucas, PH.D., Professor and Deputy Chair
DepartmentofManagementandMarketing
Fairleigh Dickinson University
Teaneck, N.J. 07666

Procedures for conducting a feedback interview include:
(a) creating a supportive climate characterized by trust, (b) emphasizing positive performance, (c) providing corroborative evidence of areas that require improvement, (d) developing faculty ownership in both problem definition and possible solutions, (e) agreeing on strategies for improvement, and (f) providing for follow-up progress reports.

Although student evaluations are being used with increasing frequency to provide input for decisions made about retention and tenure (Seldin, 1984), such evaluations are underutilized as a tool for improving teaching effectiveness. A preliminary analysis of self-report data from over 1200 department chairs in about 40 colleges and universities on the ways in which student evaluations are used to improve instructional effectiveness indicates that the most common method of providing feed-back is to place computer print-outs of teaching evaluation data in faculty mailboxes sometime during the semester following data collection (Lucas, 1989). These print-outs usually provide an analysis in which an instructor's ratings are compared with those of other faculty members teaching similar courses, with the mean scores of the entire department, or with those of the whole college. Comments by students are also included. Moreover, it is apparently assumed that since the analyses are fairly straightforward, instructors will need no support or assistance in deciding how to use such information to improve their teaching.

Little research exists to support a positive correlation between teacher evaluation and instructional improvement over time (Miller, 1988). That may be due, as Miller suggests, to the paucity of research in this area. However, it is logical to conclude that unless feedback about problems in teaching also provides personal support and ongoing assistance in correcting such difficulties, instructors may not, on their own, develop a specific plan that will bring about positive change.

If opportunities are being missed for improving instructional effectiveness, why does this happen? One likely explanation is that department chairs have been given little training in conducting successful faculty interviews, particularly the kind of interview that involves confronting inadequate or unsuccessful teaching behavior. This paper presents an overall plan and some strategies for handling a feedback interview that are designed to motivate a faculty member to work on improving his or her teaching.

To be effective in teaching requires proficiency in three areas: knowledge of one's discipline, skill in teaching methodology, and an understanding of human learning and motivation. As Edgerton (1988) has said, with respect to our fields of study most of us enter the profession of higher education standing on the shoulders of giants, but in terms of theories of learning and pedagogical knowledge, we stand on the ground.

When you have both talked about the positive elements in the individual's teaching, and in particular those strengths the individual may have taken for granted, introduce the topic of behaviors to be improved. For example a summary statement, such as, "There are many things that students value in your teaching. However, students in each of your courses seem to feel that your exams do not reflect the topics you have emphasized in class. What do you think is the reason a number of them seem to feel that way?" After the faculty member discusses his reasoning about the problem, you might paraphrase and reinforce what sounds logical to you adding other possible explanations. The idea here is not to pin the individual to the wall or accuse him but rather to help him to accept ownership (not blame students), and to view the situation as a problem to be solved.

The next step is to take a problem solving approach so that the faculty member has an opportunity to think of some alternatives he might try in class. For example, an instructor might be teaching a course in which the textbook authors provide a large pool of multiple-choice questions. In order to make the course more interesting, the professor might spend class periods giving lots of interesting case studies which involve students in active discussion. However, the class may be large, and the professor may feel that it is easier to give multiple-choice exams which can then be scanned and graded by machine. Students may then experience a discontinuity between what is discussed in class and the information they are required to produce on an exam, which may tap a large amount of memorized material but little of the reasoning they learned in class.

A department chair might best act as a catalyst here, accepting the conflict in which the instructor finds himself, yet determined that there must be an acceptable solution which can be discovered by the instructor, perhaps with some help from the department chair.

Finally, a department chair will want to work towards an agreement that a particular strategy (hopefully one developed, or contributed to by the instructor) will be attempted during the next semester. Not only should a plan be agreed upon, but implementation steps need to be worked out. Otherwise, nothing is likely to come of this interview. In addition, the chair will want to set a date for a follow-up meeting so that the faculty member can report on progress. At that time, if the plan is not working, some changes will have to be made. If things are going well, the role of the chair is easy. All that is necessary is to be very supportive and reinforcing. A comment such as, "It looks as if you have that problem solved. That should really make a big difference when your next evaluations are done."

In summary, while student evaluations of teaching are being used increasingly as a basis for personnel decisions on retention and tenure, their use as a tool for instructional development has often been neglected. Simply providing data from such evaluations to faculty is not a sufficient basis for expecting improvement in teaching. Feedback to faculty must be handled in a supportive climate in which trust has been established. A chairperson can make this interview an opportunity to reinforce positive teaching behaviors in which a faculty member is already engaged. For those skills that require improvement, a problem solving approach can be used, implementation steps developed, and a plan for reporting on progress arranged. Since such an approach follows sound psychological principles for successful behavior change, using such methodology should have a significant effect on whether student evaluations will improve a faculty member's teaching.

One important responsibility of a department chair, therefore, is to work with new faculty members before they become disenchanted with teaching, helping them by reinforcing what they are doing well in the classroom and enabling them to work out a plan for improving those areas which they are not handling successfully.

A few basic findings from the research literature about the use of student evaluations should be mentioned first. Student ratings should be used as only one aspect of a valid, comprehensive evaluation of teaching. Judgments about some aspects of teaching are better made by colleagues or a department chair; e.g., knowledge of the discipline, and appropriate emphasis in the course on significant topics.

Moreover, there is agreement in the research literature that student evaluations are valid as one aspect of teacher effectiveness. Based on an extensive review of the literature, McKeachie (1979, p. 390) concludes that student ratings are "highly valid as indices of achievement of attitudinal and motivational goals of education and are reasonably valid as indices of achievement of cognitive goals."

However, other factors are important to the appropriate use of student evaluations. One is that feedback is more accurate if it is based on data taken from several classes during a semester. To rely upon evaluations from only one class is always hazardous. "A minimum of 75 per cent of the registered students in a class must complete the form to assure credibility of the results" (Seldin, 1984, p. 137).

There is also evidence that evaluations from very small classes (under 10) and large classes (over 100) tend to be somewhat more positive than those from other classes. This apparent contradiction is explained by the fact that universities often assign their best professors and additional resources to large classes (Centra, 1979). Moreover, evidence indicates that courses in the major and electives tend to be rated more highly than required non-major courses (Centra, 1979). Such overall information is useful before taking on the task of interpreting results to a professor.

Chair preparation for a faculty interview for improving teaching, includes studying the data from several classes to pick up general trends--the aspects of teaching that the faculty member seems to be handling very well and those behaviors that need improvement. The chair must set aside enough time (30 to 45 minutes), a private place, and not allow telephone calls or other interruptions.

Professors whose teaching is not going well are often disheartened. Such individuals needs encouragement, not further criticism. A supportive climate is most likely to result in cooperation and agreement about what steps need to be taken to ensure improvement. An interview might begin with a minute or two of small talk, followed by setting the stage for the topic to be discussed. For example, "As I indicated when we made this appointment, I would like to talk with you about teaching, in particular, your student evaluations. How has your teaching been going this past semester?" After giving the faculty member a chance to speak about this, "I'd like to talk with you about those parts of your teaching that have gone particularly well. What do you feel your strengths have been?" Or, "What have you tried this semester that you feel good about?" It is important to use active listening here (that is, paraphrasing what you have heard) and to add any reinforcing comments that are appropriate. Faculty often denigrate their teaching and don't sufficiently savor their successes. It is important, however, that you do not say positive things about what are essentially areas that require improvement. Don't turn weaknesses into strengths to soften the impact.

References

Centra, J. A. (1979). Determining Faculty Effectiveness: Assessing Teaching, Research, and Service for Personnel Decisions and Improvement San Francisco: Jossey-Bass.

Edgerton, R. (1988). "Melange." The Chronicle of Higher Education, April 20.

Lucas, A. (1989). The Department Chairperson's Role in Enhancing College Teaching San Francisco: Jossey-Bass.

McKeachie. W.J. (1979). Student ratings of faculty: a reprise. Academe, 65, 384-397.

Miller, R.I. (1988). Evaluating Faculty for Promotion and Tenure. San Francisco: Jossey-Bass.

Seldin, P. (1984). Changing Practices in Faculty Evaluation: A Critical Assessment and Recommendations for Improvement. San Francisco: Jossey-Bass.

THE LEGAL REQUIREMENT FOR EMPLOYEE DEVELOPMENT: THE PUBLIC SECTOR LEADS THE WAY

Rosemarie Twomey, J.D.
Fairleigh Dickinson University
College of Business Administration
Madison Campus
Madison, New Jersey 07940

A good performance appraisal system will help both employers and employees to accomplish their respective objectives. At the same time, it will help the employer keep clear of legal entanglements in the form of employee claims of wrongful discharge and discrimination.

Employees are increasingly litigious and they have an arsenal of legal arguments at the ready to protect their rights to fair and equal treatment and job security. Some of the laws which directly impact on appraisals are Title VII of the Civil Rights Act of 1964, Executive Orders 11246 & 11375 of 1965 & 1967, the Age Discrimination in Employment Act of 1973 & 1974, the Pregnancy Discrimination Act of 1978, and the Equal Pay Act of 1963. These statutory laws are applicable in varying degrees to both the private and the public sector. In addition, the public sector must comply with constitutional law and the Civil Service Reform Act of 1978.

The above laws are directly applicable to the evaluation function of performance appraisal and their effect on appraisal is well documented in the literature. There are other laws which apply to the development function of personnel appraisal, and this article focuses on the legal parameters of this aspect, especially as experienced in the public sector.

It is well-known that employers must comply with numerous antidiscrimination laws and must not discriminate against persons with regard to their race, color, sex, religion, age, national origin, or health, and for that reason performance appraisal systems have undergone transformations to ensure that selection criteria are objective, that raters are well trained, and measurements are job-related. Not as well known to performance appraisal practitioners is the impact of a growing movement involving wrongful discharge cases which are not based on discrimination claims. In spite of the absence of legislation prohibiting dismissal except for good cause, employees are alleging wrongful dismissal based on several different legal threories including but not limited to contract breach, common law arguments in tort, and violation of public policy. For instance, employees have been successful in arguing that their dismissals were improper and illegal because there were implied contract promises of continued employment as evidenced in a policy manual or handbook, or that the employer acted in "bad faith" by not informing a recently hired employee that there was a possibility the position would soon be terminated. It is theorized here that there is a link between this surge of wrongful dismissal cases which basically protect employees' right to job security, and the right of employees to improve performance before summary discharge. The instability and unpredictability of current law in this area has led to to efforts in several states to pass laws which essentially would prohibit discharge except for "just cause". To date, Montana is the only state which has actually passed such legislation. It is posited here that passage of such a law is the first step toward establishing the right to development--the right to be given an opportunity to change before adverse action is taken against an employee. The reasoning is twofold: (1) The legislation itself establishes public policy favoring continued employment as long as an employee performs satisfactorily. Implicit in that policy is an employee's right to fairness in employment decisions. Whether or not an employee is treated fairly may depend on the resources available to that employee in his honest effort to do a good job. (2) The requirement of evidence justifying the dismissal spotlights the organization's policies and practices regarding treatment of employees. To the extent the employee can persuasively argue, "I was ready, willing, and able to learn, but I was not given direction, information, support, or training to carry out my assigned tasks", the organization may not be able to adequately defend the charges and sustain its "just cause" position.

As stated above, public employees enjoy constitutional protections unlike their counterparts in the private sector. The employee might argue for instance that the right to continued employment, absent misconduct, is a "property" right which cannot be taken away without "due process". Due process has been interpreted to include fair treatment, such as notice and just compensation. With regard to employment, due process could be expanded to mean, among other things, that employees who are diligent, but perform unsatisfactorily, are to be given whatever help is necessary to enable them to reach their employment potential.

The Civil Service Reform Act of 1978 was the first major piece of legislation to address the importance of the development function. This statute states that the performance appraisal system should put primary emphasis on the quality of an employee's work and that performance evaluations of supervisors should consider the performance of his or her subordinates. Supervisors are to use the results of the subordinates' evaluations to take one of the following courses of action: recognize and reward the employee; assist employees whose performance is unacceptable to improve; or reassign, demote, or separate the employees whose performance continues to be unacceptable. Furthermore, it specifies that adverse action should not be taken against an employee with an unacceptable performance rating until after the employee has had an adequate opportunity to improve performance. Once an employee's performance is deemed to be unacceptable, the employee is entitled to at least 30 days written advance warning of any action proposed, the expected standard, and the areas of unacceptability. (Taking adverse action within the notice period, however, is not ruled out.)

One of the primary aims of the Civil Service Reform Act was to improve the effectiveness and efficiency of the workforce by enabling the government to spot unsatisfactory performance and remove such performers expeditiously while still remaining within the dictates of the law. CSRA contains the framework for an appraisal system that would provide the government with solid legal grounds for any just cause dismissal if the development aspect of the system is fully utilized.

Much could be done to improve and build on the CSRA performance appraisal systems to enhance employee development. For example, in Utah the state government developed a policy to help underachievers upgrade their performance levels. Their experience showed that very few people need to be fired for substandard performance. The plan provides that when an employee is not performing up to standard, step-by-step corrective action is implemented. Possible corrective actions include: closer supervision with help as needed, training, referral for personal counseling, reassignment or transfer, appropriate leave, career counseling and outplacement, and constant review of performance. It is recognized that after taking corrective action and following through, it the worker still is below standard, management can fire the employee. The importance of carrying out the plan completely was borne out in the case of a procurement officer for a city in Utah who was dismissed during the "closer supervision" period. The arbitrator ordered reinstatement and back pay because evidence of assistance during that period was lacking. (White, 1985)

In addition to providing legal defenses and aiding the underachievers, it is believed that this type of approach, designed to enhance employee development, would tend to have other beneficial effects as well. When employees view the performance appraisal process as providing means for improvement, training, and development, their acceptance of the system is increased, and as they become more competent, the organization's effectiveness improves and the employee experiences confidence and job satisfaction. A 1983 survey of Iowa public employees found that roughly two-thirds of the 340 respondents found the MBO-based performance appraisal system to be of very little help with regard to either planning for or receiving needed training and development. These responses had a high correlation with the respondents' perception of diminished organizational effectiveness, responsiveness to the public, and job satisfaction--three major components measuring an organization's success. (Daley, 1987)

Providing development for employees is beneficial not only to the individual workers, but in several respects to the organization as well. It is less expensive and time consuming to help improve an employee's performance than it is to defend a just cause dismissal and have to hire and train a replacement. Performance appraisal systems which have built-in enabling mechanisms to aid employees whose performance is less than satisfactory and which attempt to get the most out of each employee's potential engender acceptance of the system in all its facets. To the extent that such a system is supported and adhered to by the entire organization, management will have the evidence it needs to successfull. defend against wrongful dismissals in regard to those cases in which continued incompetence makes dismissal for cause a necessary action.

REFERENCES

Robert N. White, "Training Today - Corrective Action - A Treatment Plan for Problem Performers", *Personnel*, Vol. 62, Feb. 1985, p. 7.

Dennis Daley, "Performance Appraisal and the Creation of Training and Development Expectations: A Weak Link in MBO-Based Appraisal Systems", *Review of Public Personnel Administration*, Vol. 8 No. 1, Fall 1987, p. 1.

MANAGERIAL BEHAVIORS AND OBJECTIVE RESULTS

Charles F. Youngberg, Ph.D.

INTRODUCTION

Of the many decisions made in an organization during the course of a year the one which employees feel is of the greatest importance to them is the annual salary decision. It is rather unfortunate, therefore, that so many organizations use approaches which are completely inadequate to the task. These include the oldest and still most popular approach -- the graphic rating scale -- which is used by 70 % of companies recently surveyed e'en though it can be shown to be immoral, illegal, demotivational and in the words of many of its victims, "just plain stupid".

Another 8% of the surveyed companies use either straight ranking, forced distribution, or a combination approach sometimes referred to as forced ranking. All three of these have been rejected by the courts when challenged because they base salary-increase decisions on employee comparisons. Of the remaining 22% of the respondents, most are using some variation of the narrative report often involving a weak MBO approach. A handful of companies have been experimenting with a variety of BARS approaches all of which appear to be unnecessary modifications of the only legally, technically and socially-acceptable performance evaluation system available today -- the Weighted Behavioral Checklist.

While the particular Weighted Behavioral Checklist reported on here was developed for the job of bank manager, any carefully-constructed WBC will have a high degree of content validity due to the fact that the scaled items are statements of the actual everyday job behaviors of the persons to be observed regardless of the particular purpose of the observations.

METHODOLOGY

There were 120 items in the original WBC consisting of 20 behavioral statements in each of 6 major job categories. The five response categories used initially were for the familiar frequency dimension -- "Never", "Rarely", "Sometimes", "Usually", anu "Always". We used the usual weights, 1 to 5, for each of the response categories respectively.

On-going research at client companies over a period of several years led to three major modifications in the format of the instrument and the method of scoring it. We found that we could get total-scale and sub-scale results of an equally high degree of reliability (upper .80's to lower .90's) using only 90 items as we had with the original 120-item version. Secondly, as a result of semantic differential research we changed the response-category weights to 10 for "Always", to 8 for "Usually", to 4 for "Sometimes", to 1 for "Rarely" and to 0 for "Never". Lastly, we devised a method for weighting each of the behaviors on the basis of how important the managers said it was to the overall successful performance of the job. The score obtained on any individual item is the product of its importance weight (obtained in advance of the period of observation) and the frequency weight (applied at the time the observations are recorded). The total score could therefore range from zero to one thousand.

Since the scaled items are weighted both vertically and horizontally we named this new approach the Dual-Weighted Behavioral Checklist. This is the first official, public mention of this new version of the WBC, its abbreviated label being the DWBC.

RESULTS

While a Weighted Behavioral Checklist has a high degree of content validity, little is known about the predictive validity of such instruments, i.e., the degree of relationship between WBC scores and independent measures of results (production, sales, profits and the like). I am here today to report on the results of one such investigation.

The population used in this study was the management staff of a medium-size bank located in northern New Jersey. Complete data on three major variables -- scores on a Dual-Weighted Behavioral Checklist, scores based on the extent to which one's objectives were attained under a formal MBO system, and scores based on salary-increase data -- were available for thirty-eight of the seventy-seven managers of the bank. The MBO results and salary-increase data were available for one and two-year intervals after the administration of the DWBC.

The results of exploratory chi-square analyses were promising enough to warrent further study. Correlational analyses revealed high positive relationships among the three variables. The strongest was between MBO achievement and salary increase ($r = .94$). This was to be expected in view of the fact that at this bank at this particular point in time salary-increase decisions were based directly on MBO results.

The relationship between the DWBC scores and salary increase ($r = .88$) was approximately equal to that between the DWBC scores and the MBO achievement measure ($r = .89$) which led us to recommend to the Human Resources Administrator at the bank that equal consideration be given to the DWBC and the MBO scores in future salary-increase decisons.

DISCUSSION

The relationship between the degree to which these managers were perceived to be following approved managerial practices on the one hand, and the extent to which they were attaining their objectives on the other, was quite high (r = .88) and statistically significant at the .01 level. This is perhaps the most important finding of the entire study since it supports quite firmly a major assumption of this investigation -- that managers who engage in those day-to-day practices which were emphasized in the bank's Management Development Program and incorporated into the DWBC do indeed achieve better results than those who do not.

The unusually high correlation of .94 between MBO results and salary increase was a reflection of a firmly-held belief on the part of top management that the merit portion of the annual salary increase should be based on the results obtained in the MBO program. Circumstances beyond the control of the individual manager are taken into consideration but once this fine tuning has been made, the results are adhered to in arriving at the size of the increase.

Of perhaps equal interest to personnel professionals everywhere, this finding provides long-needed support for the belief that effective managerial skills can be learned and that providing feedback via a weighted behavioral checklist type of instrument can lead to improved skills over a reasonably short period of time.

The results of this study also show that it is possible to design and implement an approach to measuring on-the-job performance which meets the requirements of EEO legislation and related court decisions (with respect to validity, reliability, precision and the like) for purposes of more objective and equitable personnel decisions. Although the 1978 Civil Service Reform Act applied initially to government employees, companies which do business with the government are beginning to be audited and eventually all enployers will become liable under the Act.

The main features of the Act require the development of performance evaluation systems which (a) involve job analyses wherein the critical elements of each job are clearly recorded in writing; (b) encourage employee participation in developing standards of performance; (c) inform employees of the critical requirements and standards of performance in advance of the time period for which the evaluation will apply; and (d) exclude any attempt to evaluate employees on the basis of comparison to other employees.

Closely related benefits of the establishment of a DWBC for any job category is that it can be used to uncover training needs, as a before - and - after measure for evaluating training programs, and as a criterion in the development of valid selection instruments, that is, as an acceptable way of resolving the so-called criterion problem.

While the particular DWBC described here was developed for managers, the need to develop similar instruments for other types of jobs should not be overlooked. The effective management of human resources is critical to the success of individual managers in reaching organizational objectives and a sound approach to developing and motivating the people reporting to them is of paramount importance. The DWBC can provide a giant step in that direction in view of the fact that the annual performance evaluation is the main tool used by organizations in making decisions regarding pay, promotions and related incentives.

REFERENCES

Barrett, R. S. Performance Rating. Chicago: Science Research Associates, 1966.

Carroll, S.J. and Schneier, C.E. Performance Appraisal and Review Systems. Glenview, Ill.: Scott, Foresman and Company, 1982. pp. 120ff.

Latham, G.P. and Wexley, K.N. Increasing Productivity Through Performance Appraisal. Reading, Mass.: Addison-Wesley, 1981.

Knauft, E. B. "Construction and use of weighted checklist rating scales for two industrial situations." Journal of Applied Psychology, 1948, 63-70.

Uhrbrock, R. S. "Two thousand scaled items". Personnel Psychology, 1961, 375-420.

FORMAL, INFORMAL, AND PERSONAL APPRAISALS: A CONCEPT FOR THE '90s [An Executive Summary]

Keith T. Miller, Ph.D.
Fairleigh Dickinson University

INTRODUCTION

Appraising employees within organizations can be separated into formal, informal, and personal performance appraisals. Within the confines of many organizations; formal, informal, and personal appraisals already exist. The differences in these types of appraisals are mainly due to the communicative environment within the organization. The various aspects of these appraisal systems is subsequently described.

FORMAL APPRAISALS

Formal performance appraisals are necessary when translating organizational goals into individual objectives. In many organizations, formal relationships, and therefore formal appraisals consist of two or more levels within the structure. Consensus formation for the purposes of appraising employee performance then becomes complex and must start at the top. In many organizations, approval by the representatives of the rank and file within the membership of departments is deemed necessary.

Formal performance appraisals function as follows:
----To translate organizational goals into individual objectives...
----To decide on salary increases and promotions...
----To produce documentation...
----To have a controlled forum for feedback...

Although formal performance appraisals are more widely recognized, informal appraisals may be equally as important and more widely used.

INFORMAL APPRAISALS

Every employee has a niche in the organizational community and is often informally involved in appraising supervisors, subordinates, or peers.

Relating to one another oftentimes includes informal appraisals. Informal appraisals can be planned or unplanned, and are created by supervisors or subordinates. They occur through upward, downward, diagonal, or horizontal lines of communication. Diagonal and horizontal lines of communication are particularly useful to subordinates, because they can circumvent situations that would give direct supervisors unwanted ideas or perceptions. In many instances the self-respect of subordinates can be preserved by using informal appraisals.

Informal performance appraisals function as follows:
----Rapid...
----Spontaneous...
----Oral...
----Off the Record...
----Candid...

Informal appraisals are extremely valuable as organizations move toward meeting goals, but the appearance of personal appraisals is a sure sign of organizational synergy.

PERSONAL APPRAISALS

In analyzing personal appraisals, this writer chooses to use a theory entitled, "A Theory of Democratic Learning Environments" developed by Barnes and Tidwell (1974). From this theory, four phases were chosen for the purpose of briefly describing personal relationships, therefore, also describing personal appraisals. These phases are: (1) contact, (2) consult, (3) share, and (4) choose.

Contact marks the commencement of the personal appraisal situation (Milner 1980, p.16) The contact phase of a personal appraisal occurs when one meets and relates to another person during initial interaction.

Consultation entails extending the "person-to-person" of contact into areas of more detailed inquiry or discovery. The consultation process is a particularly important phase in the personal performance appraisal process because it is at this point where the employee feels the necessity to locate and achieve certain direction with others.

Share marks the beginning of cooperative activity and growth (Barnes and Tidwell 1974, p.6). Activity where each person contributes, receives, and learns becomes the norm.

Choosing defines an additional descriptive area in personal relationships. In order to "choose", a person selects an option from two or more. To be able to choose tends to create productive appraisals (Milner 1980, p.24; Gibb 1978, pp.94-95).

Choice is the final phase of establishing the character of personal appraisals. Within, choice, individuals are free to succeed or fail.

SYNOPSIS

The fundamental problem in using the various forms of appraisal may be to find the optimum balance. It is this writer's thinking that there are untapped reservoirs of psychological and sociocratic energy in employees that could be released by more intelligent use of complete performance appraisal systems featuring formal, informal, and personal appraisals.

MAJOR REFERENCES

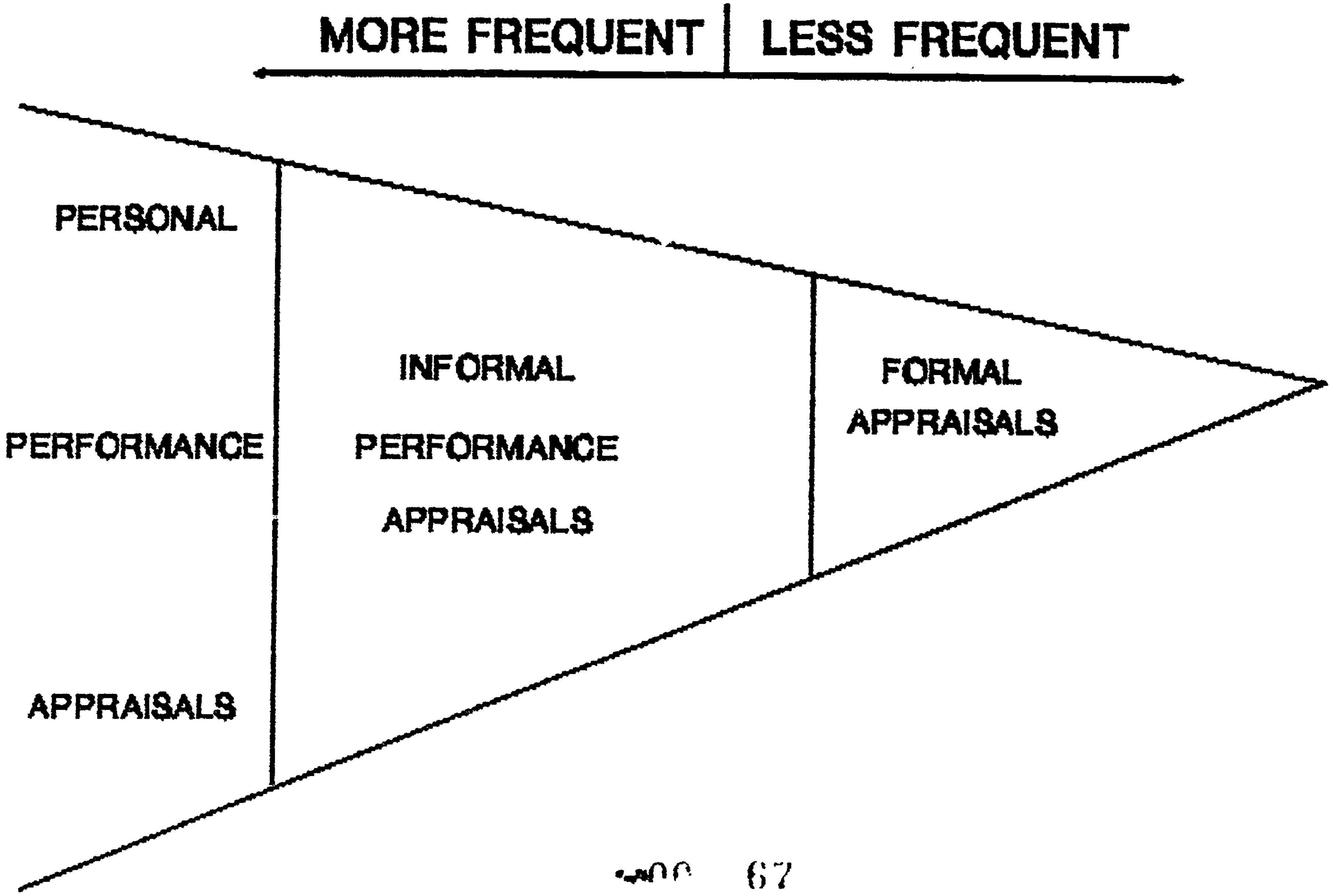
Barbery, Frazier. "Perceptions and Reported Behaviors of Participants in a Teacher Strike." Doctoral Dissertation University of Arizona, Tucson: 1980.

Barnes, William D., and Clyde D. Tidwell. "A Theory of Democratic Learning Environments." University of Arizona, Tucson: 1974.

Gibb, Jack R. Trust: A New View of Personal and Organizational Development. Los Angeles: The Build of Tutors Press; 1978.

Milner, Don R. "A study of Perceptual Patterns of Youth-in-Trouble Regarding Personal Relationships in the School and the Home." Doctoral Dissertation, University of Arizona, Tucson: 1980.

PERSONAL - INFORMAL - FORMAL PERFORMANCE APPRAISALS



MORE FREQUENT | LESS FREQUENT

PERSONAL PERFORMANCE APPRAISALS

INFORMAL PERFORMANCE APPRAISALS

FORMAL APPRAISALS

Summary of

Empirical Validity Evidence for a Task-Based Job-Component Job Analysis

by

Ronald A. Ash

School of Business

University of Kansas

Paper presented at the Annual Conference of the International Personnel Management Association Assessment Council, Orlando, Florida, June 18-22, 1989.

Job analysis is a cornerstone of effective personnel/human resource management (P/HRM). Information obtained from job analyses is used in carrying out the primary functional activities of organizational staffing, employee training and development, employee compensation, labor relations, and job design/redesign. Unfortunately, the practice of job analysis has been characterized by the use of relatively time- and labor-intensive methods for collecting job analysis data. These methods generally require job incumbents and supervisors to spend hours participating in interviews or completing long open-ended or highly structured questionnaires. The latter, in the form of task or knowledge/skill/ability (KSA) inventories, are very time-consuming and expensive to develop. The former are time-consuming and expensive to analyze. Many large organizations question the utility (costs v. benefits) of their job analysis programs, while many small organizations lack resources to mount effective job analysis programs in spite of apparent legal requirements that many P/HRM functional activities be based on job analysis data.

Job-component job analysis is a part of what several authors call job-component validity. Task-based job-component job analysis is more broad in some ways but more narrow in scope than job-component validity, and stops short of establishing empirical validity of predictors for person requirements connected to joh components. Task-based job-component job analysis involves (1) determining the task-oriented work dimensions for an occupational area, and (2) determining the knowledge, skills, and abilities (or person requirements) for each respective task dimension. This data base then serves as a foundation for development and implementation of a variety of human resources management activities and programs which can be easily and specifically tailored to particular jobs within the occupational area.

The International Personnel Management Association Assessment Council (IPMAAC) has recently conducted a task-based job-component job analysis for the occupational area of personnel assessment (Ash, 1988). In this job analysis, 15 task-based job dimensions were derived empirically for this occupational area. KSAs required for journeyman level performance of each of the 15 job dimensions were systematically identified and reliably linked to each job dimension independently. The product is a job analysis system for the personnel assessment occupation that now can be used to derive detailed ta k and KSA data for personnel assessment jobs with a minimum investment of time and effort on the part of users. Specifically, a job incumbent and/or supervisor is asked to read brief descriptions (several sentences) of the 15 task-based job dimensions, and to indicate the 1) the proportion of time spent and 2) the proportion of importance of each dimension for the target job by allocating 100 points across the 15 dimensions plus an "other activities" category. This takes less than 30 minutes per incumbent/supervisor. The proportion of importance ratings are "plugged into" the job analysis system to derive detailed task and KSA information for the target job.

Job analysis systems such as this one, once developed, appear to hold potential for reducing the time and labor required of supervisors and job incumbents relative to the more traditional practices of job analysis. However, this potential will be realized only if the job analysis data obtained from such systems are accurate and valid.

A nationwide sample of individuals working in various personnel assessment jobs completed the brief job analysis procedure described above, and also completed a 104-item KSA inventory for their respective jobs. Each respondent's ratings of relative proportion of importance for the 15 task-based job dimensions was used to determine which KSAs "should be" important in the performance of his or her job. Specifically, any dimension which received a proportion of importance rating of 10% or higher was considered important to that individual's job, and all the KSAs associated with that job dimension were categorized as "required" for that job by the job-component job analysis procedure. The remainder of the KSAs were categorized as "not required" according to the job analysis procedure. This dichotomous designation for the set of 104 KSAs was compared to that obtained directly from the respondents themselves. If the respondent rated a KSA as "moderate" (2) or "high" (3) in importance for job performance, the KSA was categorized as "required" according to the job incumbent. If the KSA was rated as "not required or needed" (0) or "low" (1) in importance for job performance, it was categorized as "not required."

Several indices of the degree of association for comparison of the KSA designations from the task-based job-component job analysis procedure with the direct designations furnished by 109 respondents were computed. On average, the job analysis procedure designated KSAs as "required" or "not required" in the same way as did the individual job incumbents 77% of the time. However, a certain amount of agreement is to be expected by chance. The proportion of agreement _after_ chance is removed is 23 to 24%, on average. This level of non-chance agreement is statistically significant at $p < .01$.

To simulate how the task-based job-component procedure might work on multiple position jobs, the proportion of time spent job dimension ratings for the 113 jobs in the data set were subjected to the SPSS-X QUICK CLUSTER procedure in order to form homogeneous job groups to serve as proxies for multiple position jobs. Six job groups were formed. Average proportion of importance data for each group was used to determine which job dimensions were considered significant, and hence, which KSAs were "required" according to the task-based job analysis procedure. Average KSA ratings were used to determine which KSAs were "required" according to job group incumbents. The same cutoff values used in the study of individuals were applied in the study of job groups. The average proportion of agreement across the six job groups is 64%. The average proportion of agreement _after_ chance is removed is 24%, again statistically significant at $p < .01$.

This study has presented empirical evidence for the validity of KSA information supplied by a task-based job-component job analysis system requiring less than 30 minutes of rating time per supervisor _or_ incumbent for the personnel assessment job under analysis.

GROUPING JOBS FOR TEST DEVELOPMENT AND VALIDATION

Julie Rheinstein, Donald E. McCauley, and Brian S. O'Leary
U.S. Office of Personnel Management
Washington, D.C.

INTRODUCTION

Developing examinations for individual occupations is always expensive and, depending on the number of hires into each occupation, may not be cost-effective. Grouping jobs on the basis of work behaviors provides one way of reducing the cost of examination development while not sacrificing test validity. This is a common approach. Barnes and O'Neill (1978) grouped jobs for examination development in the Canadian Public Service. Rosse, Borman, Campbell and Osborn (1983) clustered U.S. Army enlisted jobs into homogeneous groups according to rated job content in order to choose a representative sample of MOS's for test validation purposes.

The purpose of this study was twofold: (1) to examine three exploratory methods of grouping occupations and (2) to compare two sources of job analysis data. The present study examined factor analysis, cluster analysis, and multidimensional scaling to see if the findings of Rheinstein, McCauley, and O'Leary (1988) would be replicated. The present study also examined two sources for obtaining job analysis data--staff from a personnel research and development group and agency personnel staffing specialists.

METHOD

Data Collection. One hundred thirteen professional and administrative occupations in the civilian, federal work force were studied. Personnel research professionals and staffing specialists grouped the occupations into categories according to similarity of work behaviors. These raters were given descriptions of the 113 jobs which were taken from the federal government's Handbook of Occupational Groups and Series of Classes (1969). The job descriptions consisted of the job title and a brief narrative which summarized the major duties of the job. These job descriptions were printed on 5 x 9 cards and given to the raters for sorting. The General Schedule (GS) series numbers were not included. Raters were asked to sort the jobs according to similarities in work behaviors. No limitations were put on the number of categories each rater could generate.

Two groups completed the sort: (1) nine members from the Office of Personnel Research and Development (OPRD) at the U.S. Office of Personnel Management consisting of eight personnel research psychologists and a personnel staffing specialist (the "psychologists") and (2) seven personnel staffing specialists from seven different federal agencies (the "staffing specialists").

The categories resulting from each of the sorts were transformed into a 113 by 113 matrix for each rater wherein a one in a cell indicated that those two jobs were placed in the same category by the rater and a zero in a cell indicated that the two jobs were not placed together. The matrices thus derived were added together producing two summary matrices - one for the psychologists and one for the staffing specialists. The values in the matrices ranged from zero (no rater put the two jobs together) to nine (all psychologists put the two jobs together) or seven (all staffing specialists put the two jobs together). The matrices were analyzed using Cluster Analysis (CA), Factor Analysis (FA), and Multidimensional Scaling (MDS).

Opinions expressed in this paper are the authors' and do not necessarily represent the official policy of the U.S. Office of Personnel Management.

RESULTS

The results section will be divided into two parts. The first part will compare the three methodologies across groups of raters. The second part will compare the two groups of raters across methodologies.

Methodologies--Psychologists

In the Rheinstein et al. (1988) study, the number of jobs per cluster was relatively stable across the three solutions with the exception of the third cluster. The interrater reliability for the psychologists was found to be .53. It should be noted that inclusion of jobs in MDS clusters is more subjective than in the other two methods and that some jobs could reasonably be included in more than one MDS cluster.

The percentage of jobs in which the three solutions agreed was 83%. Percentages of agreement among the three pairs of analysis methods ranged from 80.5% to 91.2%.

The MDS solution produced 5 tightly grouped clusters and two more disparate clusters. The two disparate clusters were Cluster One and Three and of those two the more disparate was Cluster Three. MDS Cluster Three could be viewed as a combination of one or two smaller clusters and outliers.

From the MDS plot of the 113 jobs in three-dimensional space, it would appear that the large number of jobs in the CA Cluster Three was due to the forcing of all outliers into that cluster. When the differences in the FA and CA solutions were interpreted in the light of the MDS three-dimensional plots, the FA solution's disposition of the outliers was almost always supported more strongly by the MDS evidence than was that of the CA analysis. The FA tended to combine the outliers with clusters to which they were in more close proximity in three-dimensional space while the CA tended to lump them all together in one large cluster.

Staffing Specialists

In the present study, the number of jobs per cluster was less stable across the three solutions for the staffing specialists than for the psychologists. There were more noticeable differences in the number of jobs assigned to Clusters One and Five across the three methods and a greater difference in the number of jobs assigned to Cluster Three than occurred with the psychologists' data. The interrater reliability for the staffing specialists was found to be .59, almost the same as for the psychologists.

The percentage of jobs in which the three solutions agreed was 66%. Percentages of agreement among the three pairs of analysis methods ranged from 65.5% to 81.4%.

While the agreement among raters across analysis methods was lower for the staffing specialists than for the psychologists, similarities among the two sets of raters did occur. The MDS solution produced 6 tightly grouped clusters and one more disparate cluster. The disparate cluster was again Cluster Three.

Again, comparing the FA and CA solutions using the MDS plots in three-dimensional space, it was evident that the CA solution forced outliers into the one large cluster, whereas the FA solution combined the outliers with clusters to which they were in closer proximity in three-dimensional space.

Comparison of Raters

The overall agreement between the two groups of raters was good. There was a correlation of .63 between the matrices for the staffing specialists and for the psychologists.

Contrasts of the FA and CA solutions for the psychologists and the staffing specialists revealed that there was perfect agreement as to the placement of jobs into clusters among 70% of the jobs in the CA solutions and among 60% of the jobs in the FA solutions. The MDS solutions were not compared due to the more subjective nature of the inclusion of jobs in clusters.

The figure of 60% agreement between the two FA solutions is perhaps an underestimate and needs further clarification. There was one group of five jobs which were grouped together by both sets of raters in the FA solutions but which were assigned to different clusters. These five jobs had relatively large secondary loadings on the factor which was the primary factor for the other group of raters, and the difference between the two loadings was not great, indicating that the jobs could go in either cluster.

Additionally, there was a common group of 22 jobs for which there was disagreement between the two groups of raters in both the FA and CA methods. In ten of these 22 jobs, there was a relatively strong secondary loading in one group of raters on the factor which would match the primary factor for the second group of raters. For these ten jobs, the factor loadings were relatively weak on all factors. In other words, there was no clear indication of the job belonging to one cluster or another, and the difference between the primary and secondary loadings was small. These jobs could easily be put in the cluster which would agree with the other group of raters.

Thus, taking these two facts into consideration, there was approximately 79% agreement between the two groups of raters in the CA solution and 73% agreement between the two groups of raters in the FA solution.

DISCUSSION

These analyses were exploratory in nature and the clusters that emerged in this study should not be interpreted as definitive for operational purposes. The aim of this study was the comparison of the three methodolc;' s and the comparison of the groupings formed by the two sets of raters. Future studies will compare these holistic groupings to the results of a traditional task-inventory job analysis.

The three methodologies provided similar job groupings with some variations; the agreement across the three methodologies was 83% for the psychologists and 66% for the staffing specialists. Despite the difference in overall agreement across methods for the two groups of raters, the findings of Rheinstein et al. (1988) as to the utility of the different grouping methods were replicated. The variations in agreement across analysis methods appear to be due to the treatment of outliers. While each methodology gives different kinds of information, multidimensional scaling provided the best information about the outliers.

Cluster analysis produces clean clusters but gives no additional information. A job is either in or out of a particular cluster.

Factor analysis provides more information than the cluster analysis procedure. In addition to the primary factor loadings, loadings on other factors (clusters) are given. It was found that often the sizes of the primary and secondary loadings were very similar, meaning that an occupation could go in one group as well as another, or that the occupation was an outlier. Gandy (1979) also concluded that factor-analytic results were more interpretable than those from hierarchical cluster analysis when he grouped jobs for validity generalization purposes.

Multidimensional scaling gives a graphic picture of the interrelationship among the different groupings. The three-dimensional representation of the interrelationship among the occupations facilitates the placement of outliers.

The approximately 75% agreement between the two groups of raters is quite strong, especially given the fact that the raters were free to sort the jobs into as many categories as they wished and that none of the raters was familiar with all 113 jobs. The clusters found among the 113 jobs across analysis methods were very similar for both groups of raters.

The disagreements between the two groups of raters are equally interesting. Some of the disagreeme' would be explained if there were some jobs included among the 113 which do not reall) well with the other jobs. The fact that the FA and CA solutions showed disagreement between the two groups of raters in the classification of the same 22 jobs gives support to this view. It appears that both groups of raters had trouble deciding where these jobs should go.

A second explanation for the disagreements between the two groups of raters would be that the raters were classifying jobs according to different criteria. When the dimensions from the MDS solutions were examined, it was found that the first dimension was the same for both groups of raters (Program Administration vs Program Enforcement or Compliance). The viewpoints of the two groups of raters differed more on the second and third dimensions. On the second dimension, the psychologists differentiated between rule generating and rule applying occupations whereas the staffing specialists contrasted information gathering and dissemination occupations with those involved with compliance with rules. On the third dimension, the psychologists contrasted information dissemination occupations with personnel-related occupations while the staffing specialists differentiated between occupations which dealt with supplies or the environment and occupations which dealt with manpower. It would seem that the two groups of raters were looking at the jobs from similar, but not identical, viewpoints.

Another explanation of the differences could possibly be the "unit of analysis" effect found by Cornelius, Carron, and Collins (1979). While all raters were given the same narrative job descriptions, their degree of familiarity with the jobs cannot be assumed to be equal. It could be that raters' lack of familiarity with certain jobs caused them difficulty in classifying those jobs. This argument is supported by the fact that not all federal agencies have employees in all of the 113 occupations under study and that some occupations are single-agency jobs.

In summary, this study indicastes that factor analysis provides more interpretable results than does cluster analysis and that MDS used in combination with factor analysis provides greater insight into the structure of the data. The study also indicates that the amount of agreement between research-oriented and operations-oriented raters can be quite high. Comparing these small-sampled, holistic groupings to those obtained in qa full job analysis will determine their accuracy.

REFERENCES

Barnes, M. & O'Neill, B. (1978). Empirical analysis of selection test needs for 10 occupational groups in the Canadian Public Service. Paper presented to the meeting of the meeting of the Canadian Psychological Association, Ottawa, June, 1978.

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

Gandy J. (1979). Cluster analysis versus factor analysis in defining job groups. Presentation at the Conference of the Military Testing Association, San Diego, CA.

Rheinstein, J., McCauley, D.E. & O'Leary, B.S. (1988). A comparison of methodolgies for grouping large numbers of occupations. Presentation at the Conference of the Military Testing Association, Arlington, VA.

Rosse, R.L., Borman, W.C., Campbell, C.H. & Osburn, W.C. (1984). Grouping Army occupational specialities by judged similarity. Unpublished papaer, 1984.

U.S. Civil Service Commission. (1969). Handbook of occupational groups and series of classes. Washington, DC: U.S. Civil Service Commission.

Task-Ability Linkage Studies for Multiple Occupations for a Common Test

Frances S. Brogan and Cynthia C. Diane
U.S. Office of Personnel Management
Office of Personnel Research and Development
Washington, D.C. 20415

The Federal Government is currently conducting a simplification effort for selection testing for Federal occupations by combining similar occupations under a single testing procedure where feasible. The purpose of this simplification process is to increase efficiency and to reduce redundancy in the Federal testing program. As a part of this effort a project has been undertaken with Federal law enforcement occupations to determine whether such simplification is possible with these occupations. The goal of the project is to develop a single paper-and-pencil cognitive abilities test which can serve several law enforcement occupations. Hirsh, Northrop, and Schmidt's (1985) validity generralization study provides strong support for this effort.

This project has three major phases: (1) a comprehensive job analysis to determine the tasks important to the occupations, (2) a task-ability linkage to link the critical tasks to the abilities required to perform the tasks, and (3) the actual development of the test or tests to measure the abilities found to be important based on the task-ability linkage. This paper will focus on the task-ability linkage studies for the occupations, but the other phases will be touched upon to give a complete picture of the process.

Only those law enforcement occupations which were not covered by the Professional and Administrative Career Examination (PACE) were considered in the simplification process. This was a result of the continuing Luevano consent decree which affects examining for those occupations. Other occupations with less than fifty hires annually below the GS-7 level were also omitted, as well as two single-agency occupations that opted to retain their current selection procedures. The occupations finally included in the process were guards, police officers, and Deputy U.S. Marshals.

Federal agencies with more than 50 incumbents in one of the three occupations were asked to participate in the job analysis phase. Guards from three agencies, police officers from seven agencies, and Deputy U.S. Marshals from the Department of Justice took part in the job analysis.

A comprehensive task inventory was developed with the assistance of a committee of subject matter experts from five of the participating agencies. One hundred inventories were provided for administration for each of the three occupations represented within an agency. An effort was made to have adequate representation by sex, minority status, grade level (at the entry levels of GS-2 through GS-7), and central office versus field office location. Respondents were to have been on the job from six months to five years and in nonsupervisory positions.

Seven hundred and fifty-six usable inventories were returned. Of these inventories, 567 were from police officers, 148 from guards, and 41 from Deputy U.S. Marshals. The sample was fairly representative by race and national origin. Most of the respondents were at the grade 5 level and had been in their present occupation for about two years.

Methodology

The critical task lists for the task-ability linkage were derived from the responses to the "time-spent" rating in the task inventory. The respondents rated each task which they performed on the amount of time spent on that task using a scale of 1 ("very much below average") to 7 ("very much above average"). A CODAP analysis was used to calculate the mean ratings over all raters for each task within each occupation. With this method, nonperformance was given a rating of zero. This is appropriate when the importance of the task to the occupation as a whole, rather than just to those performing it, is of concern.

Although the same job analysis inventory was used with all the occupations, separate critical task lists were developed for each. This procedure assured that the important tasks for each occupation were considered in determining the abilities to be measured and is imperative to determining whether the same selection test will suffice for all the relevant occupations. In an effort to include a sufficient number of tasks to cover the abilities needed in the occupations, while keeping the performance of the linkage within a manageable framework, a decision was made to use the top 25% of the items on the "time-spent" scale for each occupation as the critical task core lists. This resulted in three core lists of 60 tasks each. In addition, to assure that any anomalies in time spent on tasks due to sex, race, or national origin be taken into consideration, tasks in the top 25% based on time spent for any sex, race, or national origin group within an occupation which were not in the original top 25% were added to the occupation's core list. The critical task lists for the occupations then varied in length with 62 tasks for Deputy U.S. Marshals, 72 for guards, and 93 for police.

In order to complete a task-ability linkage, an appropriate list of abilities must be assembled. A list of knowledges, skills, and abilities (which will be referred to henceforward as "abilities") used in a previous study at the U.S. Office of Personnel Management provided a basis for the present list. A number of the abilities, such as "persuasiveness," were elimnated because they are not easily measured. This left a list of 17 abilities. The abilities are listed in Table 1.

A linkage form was then developed for each of the three occupations. The numbered tasks were listed down the page and the abilities were listed across the page. A rating scale from 0 ("no importance") to 7 ("critical") was listed on each page of the linkage form for easy reference. Instructions were given to rate each ability on the scale of 0 to 7 for its importance to the successful performance of each task.

Incumbents who had been on the job long enough to be familiar with the tasks involved and the requirements for learning and performing the tasks satisfactorily were selected to fill out the task-ability linkage form for their respective occupations. A representative sample by agency, sex, race, and national origin was sought from each occupation. Seven guards, seven police officers, and six Deputy U.S. Marshals filled out the linkage forms.

When all the linkage forms were completed, means were calculated across tasks and raters within each occupation for each ability. These mean ratings were then examined to determine whether a common test could be used and which abilities should be measured.

Results

Twenty raters (including four females) filled out linkage forms. The sample included twelve white raters, seven black raters, and one rater from the category "Asian or Pacific Islander." Two Hispanics were included in the group. Most of the raters had worked in their present position for an average of about seven years.

The mean ability ratings for each of the occupations are given in Table 1. In order for a common test to be used for multiple occupations, the same abilities should be important to successful performance in the various occupations.

As is evident from an examination of Table 1, the three occupations are quite similar in their ordering of the importance of the various abilities. The mean ratings for "arithmetic computation," "letter writing," "spelling," and "grammar" are all below 2.5 for all three occupations, while the means for "attention to detail," "judgment," "deal with people," "general reasoning," "oral communication," "listening," and "object perception" are all above 3.5 for all three occupations. "Memory" is rated above 3.5 for two of the occupations (Deputy U.S. Marshals and police) and "fund of information" is rated above 3.5 for police and guards. "Physical stamina" for guards and "quantitative reasoning" for police are also rated above 3.5. "Reading comprehension" and "written communication" fall between 2.5 and 3.5 for all the occupations.

Discussion

In deciding which abilities should be measured by the selection test, several factors have been taken into consideration. First, the goal of the project is to develop a cognitive abilities test. Several abilities were included in the ability list, however, which are not easily measured by a paper-and-pencil test but can be tapped in an interview situation. "Judgment," "deal with people," "oral communication," and "listening" are all more easily examined in an interview setting than with a written test. The support obtained from this study for the importance of measuring these abilities has been provided to Dr. Tressie Muldrow at the U.S. Office of Personnel Management. Dr. Muldrow has incorporated the results of this project into guidelines for selection interviews for law enforcement occupations (Muldrow, in press).

Other abilities which are difficult to measure validly and reliably in a selection setting are "memory" and "fund of information." These abilities were included in the study since the literature supports their importance for the occupations. Their importance was supported in this study as well. Measures of specific types of short term memory exist, but these do not measure the type of memory needed on the job, as indicated by the "memory" and "fund of information" constructs.

Also, the ability of physical stamina was included in the list, although physical stamina cannot be measured by a pencil-and-paper selection test. This ability was primarily retained to provide information for the participating agencies who have individually developed physical performance standards.

Of the top ten abilities (which include all but one rating above 3.5), three abilities are left: "attention to detail," "general reasoning," and "object perception." These three were chosen as the abilities to be measured in the common selection test. "Quantitative reasoning," which achieved a rating above 3.5 only for the police occupation, has been subsumed under "general reasoning," because quantitative reasoning has been traditionally associated in factor-analytic research with general reasoning (see Northrop, 1989).

In summary, the second phase of the project has produced the task-ability linkage critical to the development of a single cognitive abilities test. Thus, the selection test will be composed of three item types: a name and number comparison exercise to measure "attention to detail," a perceptual speed test of comparing simple drawings to determine similarity as a test of "object perception," and a logic-based reading comprehension test as a measure of "general reasoning."

References

Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1985). Validity generalization results for law enforcement occupations. (OSP-85-2). Washington, DC: U.S. Office of Personnel Management, Office of Staffing Policy.

Muldrow, T. (in press). *An interview guide for law enforcement occupations.* Washington, DC: U.S. Office of Personnel Research and Development, Asseɔsment Services Division.

Northrop, L. C. (1989). *The factor-analytic history of selected ability constructs.* Manuscript submitted for publication.

Table 1

Law Enforcement: Task-Ability Linkage Results

	U.S. Deputy Marshals		Police		Guards	
	(N = 372)		(N = 651*)		(N = 453)	
Abilities	Mean	S.D.	Mean	S.D.	Mean	S.D.
Attention to Detail	4.34	1.51	4.73	1.62	4.44	2.16
Judgment	4.14	1.45	4.75	1.51	4.31	1.95
Deal with People	4.00	1.92	4.22	2.17	4.06	2.13
General Reasoning	3.81	1.58	4.20	1.48	3.69	1.84
Memory	3.80	1.39	4.21	1.43	3.14	2.16
Oral Communication	3.80	1.81	4.03	2.01	3.84	1.98
Listening	3.77	1.84	3.78	1.88	3.66	2.17
Object Perception	3.72	1.77	4.04	2.13	3.53	2.19
Physical Stamina	3.22	1.74	2.85	2.32	3.55	2.04
Fund of Information	3.11	1.74	3.96	1.50	3.51	2.13
Quantitative Reasoning	3.04	1.57	3.63	2.08	2.90	2.12
Written Communication	2.77	1.89	2.90	2.18	3.01	2.05
Reading Comprehension	2.70	1.73	2.65	2.11	2.87	2.21
Grammar	2.47	1.90	2.24	2.[illegible]	2.13	2.04
Letter Writing	2.26	1.89	1.44	1.77	2.27	2.05
Spelling	2.11	1.81	1.71	1.87	2.24	2.05
Arithmetic Computation	1.63	1.72	1.55	1.52	1.54	1.78

*For Quantitative Reasoning, N = 650; and for Written Communication, N = 649.

USE OF A JOB ELEMENT QUESTIONNAIRE AND INTERVIEW TO SELECT PUBLIC SAFETY COMMUNICATIONS SPECIALISTS

SUMMARY

Jeff Sherwood
King County Personnel
Seattle, WA

INTRODUCTION

The King County Department of Public Safety Communications Center provides emergency call receiving for over 500,000 citizens living in unincorporated King County, and radio dispatching for a King County police force of approximately 560 commissioned personnel. The Communications Center is staffed by 73 Communications Specialists and 8 non-commissioned supervisors, and managed by a commissioned Police Captain and Police Lieutenant. Communications Specialists perform functions other than call receiving and dispatching, such as data control, or management of information on criminal and civil warrants, stolen vehicles, and missing persons; training; interagency liaison; and technical support; although selection has always been based on what new employees are initially trained in and assigned to--primary call receiving.

The Department hired 104 Communications Specialists from mid-1979 to mid-1988. Only 46 of those 104 are still employed. 28 of the 58 who have resigned or been terminated did not last a year; 49 of the 58 did not last two years. The average tenure of those from the sample that remain employed is just under six years. The instrument used to select those 104 employees was based on a few skills and aptitudes derived from workplace observation, but did not effectively communicate realistic job expectations, such as the nature of the calls received, the level of stress involved, and the necessity of working well with a variety of people in a closed environment.

APPROACH

A job element analysis was performed using 7 job incumbents, five of them fully qualified call receivers/dispatchers, and two supervisors(Communications Specialists Supervisor are selected from among the Communications Specialists, so both supervisors had significant experience on the floor of the Communications Center, as well as experience as trainers and first-line supervisors) The job incumbents brainstormed 188 elements, statements of behaviors that would distinguish between superior and poor job applicants. These were rated, and later edited down to 35 content items, organized under six major element headings. Only one of the six elements concerned items that had formed the content of the previous selection instrument. The other five major elements had more or less been ignored in selecting Communications Specialists.

Several different approaches were considered in developing new instruments based on the job element analysis. It was decided that a job element questionnaire would be used to initially evaluate applicants, and that applicants receiving a passing rating on the questionnaire would be interviewed using a job element format. Both instruments would rely on applicants matching their significant past achievements and accomplishments with the content items identified in the job analysis. The employment list would be established ranking applicants on the basis of their questionnaire and interview ratings.

Two Communications Specialists rated the questionnaires, and four Communications Specialists served as raters in the interview process. Training and specific instructions on the use of the instruments was provided to all the raters. The interview raters role played through three interview prior to actually rating applicants. All the raters were provided with an anchored rating scale, but how applicants specific achievements and experiences were rated was left up to the expert knowledge of the raters. For both the questionnaire and the interview, major elements were rated, from specific information gathered from the applicant's addressing of the 35 content items. 95% consensus was required of the raters and absolute consensus was required on passing or failing applicants.

CONCLUSIONS

Questionnaire and interview ratings were correlated with applicant self-ratings. Both correlations were significant at the .01 level (questionnaire r = .32, n = 133; interview r = .51, n = 53). The Department hired 14 people off the resulting employment list since June of 1988, and anecdotally has been very pleased with the results, and their ability to participate meaningfully in the selection process. There has been some turnover of these hires in the past 12 months, but none of them has left because of poor performance or mistaken notions about the work. A more thorough study will take place when there is a larger sample and long-term performance data is collected.

One of the most important side-effects of the selection process has been to isolate the factors outside of the control of the examination: low entry-level salary, slow salary growth, intermittent availability of advanced training.

For an ongoing recruitment and selection process, the job element analysis has been further edited, primarily through J-coefficient analysis, and the interview material has been revised based on feedback from raters. In the future, the questionnaire may be used for screening purposes only, if the data continues to indicate that the job element interview is more predictive of job success.

COMMUNICATIONS SPECIALIST

ELEMENT I: ABILITY TO HANDLE EMERGENCY SITUATIONS (126)

1. Immediately recognize serious situations. (71)
2. Remain calm. (19)
3. Reassure and calm other people. (80,128)
4. Control a conversation to identify and obtain necessary information. (15,25)
5. Quickly make appropriate decisions to resolve situations.

ELEMENT II: ABILITY TO DEAL WITH UNPLEASANT OR UNCOMFORTABLE SITUATIONS (107)

6. Effectively deal with emotionally distressed or anxious callers. (114)
7. Effectively deal with graphic sexual or violent situations. (112)
8. Effectively deal with child abuse or molestation. (111)
9. Effectively deal with domestic violence. (109)
10. Effectively deal with suicidal behavior. (108)
11. Effectively deal with potentially fatal medical problems. (125)

ELEMENT III: ABILITY TO WORK ON YOUR OWN WITHOUT DIRECT SUPERVISION (49)

12. Make independent decisions exercising best judgment and accept responsibility for them. (106,174)
13. Know when to refer situations to a higher authority. (172)
14. Prioritize work assignments. (177)
15. Monitor own work performance. (178)
16. Deal with problems which are not set in definite terms. (184)
17. Recognize and try alternative methods for solving problems. (3,140)

ELEMENT IV: ABILITY TO COMMUNICATE EFFECTIVELY AND WORK IN A TEAM SETTING (13,55)

18. Speak clearly and maintain voice control. (12,97,98,117,118,119,120)
19. Accurately transfer information to other parties. (26,150)
20. Give clear and accurate instructions. (32,33)
21. Listen to people. (105)
22. Work and communicate effectively with fellow employees, supervisors and staff of other organizations. (40,45)

ELEMENT V: ABILITY TO HANDLE STRESS RELATED TO THE JOB (133)

23. Handle complaints on a daily basis. (132)
24. Effectively handle complaints from angry, dissatisfied callers. (82)
25. Maintain concentration at all times in spite of varying workload. (74)
26. Adapt to changing work situations. (142)
27. Tolerate trivial irritations and interruptions. (36)
28. Keep track of and handle several events or tasks that occur simultaneously. (8,66)

ELEMENT VI: ABILITY TO LEARN AND APPLY NEW SKILLS AND KNOWLEDGE (83)

29. Follow oral and written directions. (10,124)
30. Learn and adapt to new equipment and procedures. (95,96)
31. Read source materials on rules, regulations or policies to obtain or maintain knowledge applicable to job responsibilities. (86,90,166)
32. Effectively use available resources. (93)
33. Use map books to locate addresses. (89)
34. Work rapidly and accurately with names, numbers and codes and/or symbols. (73,121,122,158)
35. Operate a variety of controls, including a typewriter keyboard. (4,7,145,160)

ASSESSING CREATIVITY IN A CONTENT-VALID EXAMINATION

Teresa F. Doyle, Ph.D. and Thung-Rung Lin, Ph.D.
Los Angeles Unified School District

SUMMARY

Dominated by the 1978 Uniform Guidelines on Employee Selection Procedures, the trend in personnel selection, especially in the public sector, has been a move decidedly away from the assessment of personality factors or traits and toward the assessment of job-related behaviors. Such a focus in selection is intended to produce more objective evaluations of prospective employees with respect to actual work expectations and, conversely, to discourage selections made on the basis of factors unrelated to an applicant's ability to perform on the job. As a consequence, the demonstration of a selection instrument's content validity, its job-relatedness and its ability to accurately differentiate among candidates' past relative success in performing work-related behaviors, has been the key to defending against charges of bias in personnel selection.

The development of content-valid examinations begins with a thorough job analysis in which the analyst attempts to determine what work behaviors.are intrinsic to successful performance on the job. Input is obtained from bcch supervisors and incumbent employees to make the determinations o: what characterizes exceptional, acceptable and unacceptable work behavior. Nevertheless, when supervisors and incumbents alike are asked to describe the ideal employee for job analysis purposes, responses frequently refer to personal attributes rather than proficiency with respect to work-related behavior. The success of the superior incumbent is often ascribed to such things as "dependability," "good attitude," "good judgment" or "conscientiousness" as opposed to technical skill. When traits are determined to be of sufficient importance to performance on the job that they must be incorporated into the selection process, they are frequently assessed in an interview format. Under these circumstances, the traits being "measured" may be loosely defined and it is left to the interviewer to categorize candidates as exceptional, acceptable or poor according to the interviewer's own conceptualization of the trait construct. Even in those cases in which the job analyst carefully ties behavioral anchors to categorical descriptions, those descriptions may be more the product of the analyst's personal construct conceptualization than that derived from a review of the psychological literature. The challenge for the personnel selection specialist is to develop an appropriate instrument for assessing traits in a content-valid examination that is consistent with empirical findings in the topic area.

The current presentation was prompted by trait-anchored descriptions of superior incumbents. Specifically, a job analysis was conducted for the purposes of designing an instrument for the selection of "Placement Coordinators," individuals who develop permanent employment opportunities for disabled high school students. Individual interviews were conducted with supervisors and incumbents in which they were asked to describe the methods used by Placement Coordinators to find jobs for disabled youth. Unfortunate from the job analyst's perspective, supervisors and incumbents agreed that specific knowledges, skills and abilities were far less important than was a particular personal characteristic in determining success on the job as a Placement Coordinator. In fact, the interviews revealed that successful Placement Coordinators shared little in common with respect to the actual behaviors employed to find jobs for students; the universal common denominator was creativity. In order to be successful, a Placement Coordinator had to be creative. Thus, our task was to construct a behaviorally-anchored selection instrument to provide an assessment of creativity in candidates applying for Placement Coordinator positions.

A literature review on the topic of creativity and its assessment was conducted as a starting point, and a fundamental problem was soon revealed. Though "creativity" has been a subject of investigation since nearly the inception of psychological research, little consensus exists with respect to its operational definition. Creativity has been variously described as a form of problem-solving ability (Cattell, 1971), the ability to generate multiple solutions to problems, also called "divergent thinking" (Guilford, 1950), the ability to recognize possibilities (Tyler, 1978), an ability to synthesize "new" problems from apparently disparate pieces of information or fragmentary problems so as to formulate comprehensive solutions (called "problem-finding" ability) (Mackworth, 1965; Getzels and Csikszentmihalyi, 1976; Kasperson, 1978; Glover, 1979), and the ability to produce novel solutions to problems (see Barron and Harrington, 1981, for a review). There was considerable agreement, however, that creativity was not so much a pervasive personality characteristic as it was contextual; that is, individuals were described as creative within their respective fields (eg., Barron and Harrington, 1981; Gruber, 1988). Our applicants, therefore, were evaluated according to dimensions consistently described as aspects of creativity, or factors strongly related to creativity, by assessing their abilities to provide solutions to job-related problems.

The final selection instrument was a paper and pencil exercise that presented candidates with various problems encountered by Placement Coordinators in their work and, after a study period, required the applicants to provide solutions. Candidates were asked, for example, to list techniques used to generate job openings for disabled students. Applicants were also presented with a biographical data sheet describing a fictitious disabled student and asked to determine what further information they needed to solicit from the student and the student's parents and teachers in order to begin a job search. Their solutions, then, were evaluated for novelty (as indicated by the number of times the same solution was proposed by different testees), diversity (as indicated by the number of different solutions provided for a single problem), problem-finding ability (the candidate's ability to recognize, collect and utilize pertinent information) and problem-solving ability (the candidate's ability to propose workable solutions to problems encountered on the job). Supervisors of incumbent Placement Coordinators, who had no prior knowledge of any of the individual candidate's capabilities, served as raters. Twenty-three candidates were evaluated in the selection process.

The selection instrument produced strong differentiation among the individual candidates: four candidates were rated "excellent," twelve as "good," four as "acceptable," and three candidates were rated "unacceptable." In addition, there was a degree of correspondence among raters with respect to the assessment of individual candidates, with nearly unanimous agreement as to descriptive score and no greater variation than adjacent descriptive scores. There were no pass/fail differences. Finally, the Special Education Department, who had requested the examination, was well-satisfied with the resultant rankings of candidates.

As a continuation of this project, the collection of criterion validity data from incumbent Placement Coordinators is currently underway to determine whether, in fact, this instrument has predictive validity for performance on the job.

REFERENCES

1. Barron, F. and Harrington, D.M. (1981) Creativity, intelligence and personality. Annual Review of Psychology, 32: 439-476.

2. Cattell, R.B. (1971) Abilities: Their Structure, Growth, and Action. Boston, MA: Houghton Mifflin.

3. Getzels, J.W. and Csikszentmihalyi, M. (1976) The Creative Vision: A Longitudinal of Problem Finding in Art. New York, NY: Wiley.

4. Glover, J.A. (1972) Levels of questions asked in interview and reading sessions by creative and relatively noncreative college students. Journal of Genetic Psychology, 135:103-108.

5. Gruber, H.E. (1988) The evolving systems approach to creative work. Creativity Research Journal, 1:27-51.

6. Guilford, J.P. (1950) Creativity. American Psychologist, 14: 469-479.

7. Kasperson, C.J. (1978) Psychology of the scientists: XXXVII. Scientific creativity: A relationship with information channels. Psychological Reports, 42:691-694.

8. Mackworth, N.H. (1965) Originality. American Psychologist, 20: 51-66.

9. Tyler, L.E. (1978) Individuality. San Francisco, CA: Jossey-Bass.

Content Validity with the Consent of Justice: A Collaborative Methodology

Doris M. Maye, State of Georgia
Irwin L. Goldstein, University of Maryland

This paper looks at content validity step by step as used in the State of Georgia for the job class Radio Operator. It highlights aspects of a recent effort conducted under a U.S. Department of Justice consent decree wherein a Justice expert, a State expert and State staff work hand-in-hand to devise and implement job analysis that is state-of-the-art and fully comporting with current professional and legal principles governing content validity.

The job analysis procedure consisted of seven major steps:

- Background Job Data Research
- Site Visits and Job Observation
- Task Identification and Development
- KSA Identification and Development
- Job Analysis Questionnaire Development and Administration
- Questionnaire Data Analysis
- Task/KSA Linkage

Key concepts guiding the project were the systematic gathering of comprehensive work behaviors -- an accurate "shapshot" of the job as it currently exists -- and the identification of discrete worker characteristics needed for adequate job performance at entry. The process involved:

- a methodical determination and sampling of how the job varies;
- the use of SMEs for the generation of a complete set of discrete tasks associated with job duty categories;
- the identification by SMEs of knowledges, skills, and abilities (KSAs) used per task;
- scaled data ratings by SMEs on task importance and KSA importance;
- scaled data ratings by SMEs on tasks and KSAs needed at entry;
- scaled data ratings by SMEs on the relationship between tasks and KSAs; and
- a priori setting of scale value criteria for task and KSA inclusion.

Of particular note is the manner in which each of the components of the process was implemented. To be sure, most content validity projects would include an outline similar to the one above. The paper uses an earlier Georgia Radio Operator (RO) project contrasted with the current consent decree project to illustrate some of the subtleties of the how versus the what in adequate content validity methodology. Among the more pertinent points raised are the following:

1. The project must be sure to describe the job as it currently exists. This begins with job use demographics such as number of positions, organization(s) found in, class series structure, etc. but also includes comprehensive focus on factors on which the job varies. Typical factors such as rural vs. urban work settings are often identified. The key "how" here is the depth of probing to ensure that comprehensive factors are recognized and used as the

sampling grid. The consent decree RO project documented eight variation factors:

- Rural vs. Urban
- Atlanta vs. Non-Atlanta
- Interstate vs. Non-Interstate
- Shift 1 vs. Shift 2 vs. Shift 3
- High # of Accidents vs. Low # of Accidents
- High # of Telephone Calls vs. Low # of Calls
- High # of Radio Calls vs. Low # of Calls
- High Clerical Workload vs. Low Clerical

2. Site visits, using the "varying factors" grid, are essential. It is necessary for the researcher(s) to see and understand the job environment and work activities in order to proceed in the remaining project steps with enough contextual frame of reference to ask the right questions, question the logic of information and data, etc. In the consent decree RO project there were eight site visits, chosen to sample the eight factors of variation. The number of visits is not the significant issue. The crux of the matter is to ensure comprehensive observation. This might be accomplished in some job classes with one or two visits, while others might be incomplete with twelve. Site visits also offer the opportunity to talk to incumbents and supervisors and to identify and collect examples of materials used on the job.

3. The most comprehensive and defensible way to obtain task and KSA data is from appropriately chosen SMEs under appropriate guidance in an appropriately systematic manner. The process used in the consent decree RO project used groups of SMEs who were representative on the varying factors grid, with representation by race, sex, and age, minimum tenure of six months, and the "request" that they be above average performers. Four formal task generation panel sessions were conducted using 16% of the population of ROs representing 63% of the Georgia State Patrol posts. The groups were asked to focus on how the job varies and to brainstorm the "big" categories -- i.e., the major duty categories -- first, then each of these was taken separately as the catalyst for tasks. The design called for the group leader to interact with the SMEs throughout and to record the information on a flip chart. By the fourth session, no "new" tasks were emerging; hence the decision was reached that comprehensive initial job behavior task information had been obtained.

The initial tasks were edited and rationally clustered in order to organize the task list and to provide a process to insure task statement clarity and completeness. Cluster definitions and task statement wordings were edited and tasks were taxonomized independently until they were consistently identified with their clusters by the researchers. The process resulted in nine task cluster categories and 64 tasks.

These tasks and task clusters were then used as the catalyst for KSA Generation Panel Sessions that followed a brainstorming/recording format similar to the task sessions. Seven formal KSA generation/review panels, using over 60% of the direct supervisors of ROs, followed by KSA editing, clustering, and taxonomizing by the researchers, resulted in 21 knowledge statements grouped into three clusters and 37 skill and ability statements grouped into nine clusters.

4. The focus and level of specificity of the task and KSA statements are critical. Tasks must describe a discrete activity, denoting what is being done, to whom, why and how. KSAs must deal with fundamental mental, physical or affective capabilities. Care must be taken to avoid a KSA statement that is nothing more than a restated task. While the absolute number of task and KSA statements is not in itself informative, it can be useful as a "red flag" indicating problems with the level of specificity of the statements. For other than the most basic and uni-dimensional jobs, less than 20 statements would likely indicate that the individual statements are not discrete enough, while over 100 statements for a single job class might raise the question of too much specificity. The number of statements will be variable and highly dependent upon the nature of the job class but should be driven, first, by completeness in capturing a "shapshot" of the job and defining needed worker characteristics and, second, by the meaningfulness of the statements to the SMEs for valid and reliable scaled data input. Following are several task and KSA statements from the consent decree RO project:

    - Responds verbally to citizens' requests concerning items such as accident reports, drivers' license information, directions, DUI and driver improvement schools, as a public service.
    - Provides emergency information, such as weather and road hazards, to the public and other non-law enforcement agencies in order to provide for public safety.
    - Searches maps for geographical information in order to respond to requests from general public, troopers, and other law enforcement personnel.
    - Logs trooper activities in direct response to radio and telephone communications for purpose of record-keeping and ensuring Trooper safety.
    - Knowledge of radio codes and call signs of law enforcement agencies in the area.
    - Ability to organize incoming information for verbal transmission on radio or telephone.
    - Ability to evaluate and interpret written sources of information, such as computer output.
    - Skill in writing numbers and letters legibly.
    - Ability to give directions in terms of compass points in the post area.
    - Ability to give geographical directions over radio and telephone accurately.
    - Ability to perform many different tasks at once.

Other aspects of the project are also discussed in the paper, such as the development and scaling of the job analysis questionnaire, the task/KSA linkage process and the criteria for task and KSA inclusion. For further information, persons should contact the principal author.

DEVELOPMENT OF A SELECTION PROCEDURE FOR TOLL COLLECTOR

Lance W. Seberhagen, Ph.D.
Seberhagen & Associates*

Objective

The objective of the project was to develop a new selection process for entry-level Toll Collector in a state transportation agency in the Mid-Atlantic region of the United States. The agency has seven toll facilities spread across the state that include bridges, tunnels, and highways.

Job Analysis

I used a multf-method job analysis approach to obtain an accurate and complete picture of job duties and worker requirements. The job analysis included: (1) literature review, (2) review of internal documents, (3) interviews with supervisors, (4) interviews with job incumbents, (5) direct observation, (6) questionnaire survey of supervisors, (7) questionnaire survey of job incumbents, and (8) management review. Exhibit 1 shows the basic results of the job analysis.

Selection Procedures

The total selection process for Toll Collector includes the following components:

1. Minimum qualifications (MQs)	Pass/Fail
2. Training and experience Rating (T&E)	Ranking
3. Written tests (TC Ability Test, Service Questionnaire)	Ranking
4. Oral exam	Ranking
5. Personnel Selection Inventory (PSI-3)	Pass/Fail
6. Medical exam	Pass/Fail
7. Background investigation (by police)	Pass/Fail
8. Classroom training	Pass/Fail
9. Probation period (6 mo., with option to 12 mo.)	Pass/Fail

Applicant Flow

All applicants must complete a special Toll Collector application blank (see Exhibit 2). This application blank is designed to assess minimum qualifications (MQs) and to rate training and experience (T&E). Only applicants who pass the MQs are rated on T&E. Applicants who pass the MQs and score well on the T&E are then invited to the written tests (i.e., TC Ability Test, Service Questionnaire). Applicants who pass the written tests are then ranked on a preliminary eligible list, according to their combined score from the T&E and written tests. Separate eligible lists are made for each region of the state, based on applicant geographical availability, as indicated on the application blank.

* 9021 Trailridge Court, Vienna, VA 22182. Tel. (703) 790-0796.

Each toll facility conducts its own oral exams. When a position becomes available at a toll facility, the toll facility conducts oral exams for the top five candidates on the regional eligibility list. If there is more than one vacancy, the toll facility invites two additional candidates from the eligible list for each additional vacancy. Each toll facility makes its preliminary hiring decision on the basis of the oral exam score alone. For purposes of selection, all candidates invited to the oral exam are considered equal, and the oral exam is used, in effect, as a "tie-breaker."

After the toll facility has made its preliminary hiring decision, the selected candidate must pass the rest of the selection process on a non-competitive basis before achieving regular status as a Toll Collector. The six remaining steps include the integrity test, medical exam, background investigation, classroom training, on-the-job training, and probation period.

Validation

Content validation provided the basis for the development of the MQs, T&E rating, TC Ability Test, and oral exam. Criterion-related validation through test transportability provided the basis for the use of two commercial tests: the Service Questionnaire and Personnel Selection Inventory. The state agency developed the remaining parts of the selection process through appropriate methods.

Adverse Impact

After about one year of use and assessment of several thousand applicants, the state has found no adverse impact, as defined by the 80% Rule and statistical significance, in any selection procedure or in the overall selection process.

Discussion

The basic minimum qualifications for this Toll Collector are simply "8th Grade and no experience," but the job is much more difficult than one might guess from the MQs. The job has many undesirable working conditions that are not obvious to applicants and which require a special type of person. In the past, many new Toll Collectors quit the job after a short time. At least one new Toll Collector walked off the job in the middle of her first day on the job! Thus, a customized application blank (see Exhibit 2) was designed just for Toll Collector to inform applicants about these undesirable conditions and to screen out anyone who was not willing to work under those conditions. This application blank measures personality in a content valid selection procedure by asking very specific job-related questions in a directly observable format (e.g., "Are you willing to wear a uniform to work?" "Are you willing to work in a job that exposes you to frequent verbal abuse...?")

Other notable selection procedures are the TC Ability Test and the Service Questionnaire. The former is a custom test that measures the ability to (1) classify vehicles under a simplified version of toll schedules used on the job, (2) make change, and (3) perform general arithmetic. The latter is a commercial test that measures service-orientation, with subscales for (1) energy, (2) acceptance of authority, (3) sociability, (4) friendliness, and (5) emotional stability.

Exhibit 1. RESULTS OF JOB ANALYSIS

JOB ANALYSIS: TOLL COLLECTOR I   March 15, 1988

JOB TASKS (Ranked by Importance within Task Groups)	Importance* Mean	Importance* SD	Frequency** Mean	Frequency** SD
REPORTS TO WORK				
1 Inserts ID key to log onto toll system before collecting tolls.	1.00	0.00	1.87	0.34
2 Arrives at least 15 minutes before assigned shift.	1.00	0.00	2.00	0.00
3 Certifies correct amount in change fund at start of shift.	1.00	0.00	2.00	0.00
4 Receives and counts change fund.	1.00	0.00	2.00	0.00
5 Reports to assigned lane 5 minutes before start of shift.	1.07	0.25	2.00	0.00
6 Reports to shift supervisor.	1.07	0.26	2.00	0.00
7 Maintains neat and clean appearance.	1.13	0.34	2.00	0.00
8 Receives lane assignment.	1.13	0.34	2.00	0.00
9 Complies with MdTA standards for grooming and dress.	1.13	0.50	2.00	0.00
10 Wears proper uniform.	1.13	0.34	2.00	0.00
11 Signs daily timesheet	1.40	0.61	2.00	0.00
CLASSIFIES VEHICLES				
12 Sees if vehicle is a semi-trailer.	1.00	0.00	1.33	0.47
13 Sees if vehicle is modified for commercial use.	1.00	0.00	1.33	0.47
14 Sees if vehicle is a light panel truck.	1.00	0.00	1.33	0.47
15 Sees if vehicle is a light pickup truck.	1.00	0.00	1.33	0.47
16 Sees if vehicle is a tractor.	1.00	0.00	1.38	0.49
17 Sees if vehicle is a truck.	1.00	0.00	1.38	0.49
18 Sees if vehicle is a bus.	1.00	0.00	1.38	0.49
19 Sees if vehicle is a passenger car.	1.00	0.00	1.36	0.49
20 Sees if vehicle is a trailer.	1.00	0.00	1.38	0.49
21 Sees if ticket is valid for toll facility where used.	1.00	0.00	1.40	0.49
22 Sees if ticket is valid for class of vehicle.	1.00	0.00	1.40	0.49
23 Counts number of axles on each vehicle.	1.00	0.00	1.43	0.49
24 Observes vehicles coming into lane of toll booth.	1.00	0.00	1.43	0.49
25 Sees if vehicle is recreational.	1.00	0.00	1.43	0.49
26 Classifies vehicle according to rules for toll facility.	1.00	0.00	1.47	0.50
27 Calls Toll Sgt when in doubt about proper classification.	1.00	0.00	1.87	0.62
28 Sees if payment is by cash or ticket.	1.07	0.25	1.40	0.49
29 Sees if ticket has not yet reached expiration date.	1.07	0.25	1.47	0.50
30 Sees if vehicle is unusually wide.	1.33	1.01	1.38	0.49
31 Sees if vehicle is unusually long.	1.53	1.20	1.42	0.49
32 Sees if vehicle is farm equipment.	1.53	1.36	1.42	0.49
33 Sees if vehicle has free pass (e.g., funerals).	2.00	N/A	1.00	N/A
34 Sees if vehicle has charge slip (e.g., military convoy).	2.00	N/A	1.00	N/A
35 Sees if vehicle is unusually heavy.	2.00	1.73	1.33	0.47
36 Sees if vehicle is carrying hazardous cargo.	2.00	1.60	1.42	0.49
37 Sees if ticket is valid for day of week used.	2.38	1.82	1.58	0.86
38 Sees if ticket is valid for time of day used.	2.46	1.78	1.36	0.48
39 Sees if vehicle has proper commuter decal.	2.63	1.87	2.17	1.34
40 Sees if vehicle has a Maryland license plate.	3.50	1.69	1.50	0.50

SD = Standard Deviation

* 5-point scale of Importance, where 1 = Very High and 5 = Ve Low.

** 5-point scale of Frequency, where 1 = Hourly and 5 = Yea..

JOB ANALYSIS: TOLL COLLECTOR I   March 15, 1988

JOB TASKS (Ranked by Importance within Task Groups)	Importance* Mean	Importance* SD	Frequency** Mean	Frequency** SD
COLLECTS TOLLS				
41 Receives cash in payment of toll.	1.00	N/A	1.00	N/A
42 Presses classification button when driver gives toll.	1.00	0.00	1.36	0.48
43 Announces value of bill received and amount of toll due.	1.00	0.00	1.36	0.48
44 Records toll payment while vehicle is at booth.	1.00	0.00	1.40	0.49
45 Counts out change in view of driver.	1.00	0.00	1.40	0.49
46 Determines proper toll based on vehicle classification.	1.00	0.00	1.40	0.49
47 Acknowledges payment with "Thank you."	1.00	0.00	1.47	0.50
48 Receives commuter ticket in payment of toll.	1.00	0.00	1.50	0.50
49 Records unusual occurrences during shift.	1.00	N/A	2.00	N/A
50 Calls Toll Sgt if wrong change given.	1.00	N/A	2.00	N/A
51 Calls Toll Sgt if driver disputes proper toll.	1.00	0.00	2.07	0.77
52 Gives receipt to driver upon request.	1.07	0.25	1.40	0.49
53 Records unusual vehicles.	1.07	0.25	1.73	0.44
54 Calls Toll Sgt if driver fails to pay proper toll.	1.07	0.26	1.93	0.57
55 Calls Toll Sgt if driver cannot pay proper toll.	1.08	0.27	1.87	0.62
56 Uses largest denomination of bill available to make change.	1.13	0.50	1.40	0.49
57 Calls Toll Sgt if driver leaves without taking change.	1.13	0.34	2.00	0.52
58 Tells driver what proper toll is for vehicle.	1.14	0.35	1.40	0.49
59 Records correction needed if vehicle was misclassified.	1.14	0.35	1.79	0.41
60 Retrieves dropped tolls in courteous manner.	1.14	0.35	1.80	0.40
61 Explains toll to driver upon request.	1.15	0.36	1.86	0.52
62 Cancels commuter ticket immediately with punch.	1.20	0.40	1.40	0.49
63 Accepts but does not give pennies as change.	1.27	0.57	1.60	0.49
64 Notes vehicle license and description if wrong change given.	1.27	0.44	1.93	0.44
GIVES ASSISTANCE				
65 Calls Toll Sgt to report damage to state property by drivers.	1.07	0.25	2.14	0.64
66 Calls Toll Sgt to relay problems reported on highway.	1.13	0.34	1.80	0.40
67 Calls Toll Sgt if vehicle needs escort on bridge or tunnel.	1.14	0.52	1.86	0.52
68 Calls Toll Sgt to report drunk drivers.	1.27	0.57	2.00	0.52
69 Calls Toll Sgt to report illegal or suspicious activity.	1.27	0.57	2.13	0.72
70 Gives directions for nearby travel routes upon request.	1.40	0.49	1.80	0.40
71 Testifies in court about events witnessed at toll facility.	1.40	0.61	3.29	1.72
72 Calls Toll Sgt to close lane if emergency on bridge or tunnel.	1.93	1.50	2.10	0.54

SD = Standard Deviation.

* 5-point scale of Importance, where 1 = Very High and 5 = Very Low.

** 5-point scale of Frequency, where 1 = Hourly and 5 = Yearly.

Exhibit 1. (cont.)

JOB ANALYSIS: TOLL COLLECTOR I March 15, 1988

JOB TASKS (Ranked by Importance within Task Groups)	Importance*		Frequency**	
	Mean	SD	Mean	SD
ACCOUNTS FOR MONEY AND TICKETS				
73 Keeps running count of money and tickets collected.	1.00	0.00	1.53	0.50
74 Counts money and tickets collected on shift.	1.00	0.00	1.80	0.40
75 Removes identification key to log off toll system.	1.00	0.00	1.87	0.34
76 Replenishes change fund during shift if needed.	1.00	0.00	1.87	0.34
77 Writes deposit slip for money and tickets collected on shift.	1.00	0.00	1.92	0.28
78 Writes report of money and tickets collected on shift.	1.00	0.00	1.93	0.26
79 Certifies correct amount in change fund at end of shift.	1.00	0.00	1.93	0.25
80 Counts money in change fund at end of shift.	1.00	0.00	1.93	0.25
81 Replenishes and balances change fund at end of shift.	1.00	0.00	1.93	0.25
82 Deposits money and tickets at toll facility.	1.00	0.00	1.93	0.25
83 Checks out with shift supervisor.	1.07	0.26	2.00	0.00
84 Keeps running count of money in change fund.	1.13	0.50	1.87	0.34
85 Secures toll booth.	1.17	0.37	2.00	0.00
86 Secures toll facility.	2.17	1.46	1.80	0.40
87 Pays difference if shortage of money is found.	2.47	1.44	2.78	1.40

SD : Standard Deviation.

* 5-point scale of Importance, where 1 : Very High and 5 : Very Low.

** 5-point scale of Frequency, where 1 : Hourly and 5 : Yearly.

JOB ANALYSIS: TOLL COLLECTOR I March 15, 1988

WORKER CHARACTERISTICS (Ranked by Importance within KSAO Groups)	Importance*		When Developed**	
	Mean	SD	Mean	SD
KNOWLEDGE OF...				
1 Different types of motor vehicles.	1.57	0.73	1.80 A	0.40
2 MDTA toll schedule.	1.64	1.11	1.67 A	0.47
3 MDTA vehicle classification system.	1.71	1.10	1.93 A	0.25
4 MDTA policies and procedures.	1.86	1.25	1.67 A	0.47
5 Counterfeit money.	2.31	1.32	1.69 A	0.46
SKILL IN OPERATION OF...				
6 Classification console.	1.57	1.12	1.93 A	0.25
7 Intercom.	1.92	1.14	1.93 A	0.25
8 Toll booth (e.g., lights, heating, air conditioning)	1.93	1.16	1.93 A	0.25
9 Calculator.	2.13	1.36	1.50 A	0.50
10 Money counting machine.	2.14	1.12	1.86 A	0.35
ABILITY TO...				
11 Get to work on time every day.	1.00	0.00	1.40	0.49
12 Count money.	1.07	0.25	1.00	0.00
13 Provide courteous service.	1.07	0.25	1.27	0.44
14 Complete simple forms and records.	1.07	0.26	1.21	0.41
15 Take abuse from public.	1.07	0.26	1.54 A	0.50
16 Perform general arithmetic.	1.13	0.34	1.00	0.00
17 Present a neat and clean appearance.	1.13	0.34	1.00	0.00
18 Read routine reports and correspondence.	1.14	0.35	1.21	0.41
19 Follow oral and written directions.	1.20	0.40	1.00	0.00
20 Exercise manual dexterity.	1.21	0.41	1.15	0.36
21 Write routine reports and correspondence.	1.23	0.42	1.38	0.49
22 Speak clearly and audibly.	1.33	0.60	1.00	0.00
23 Write legibly.	1.33	0.60	1.07	0.25
24 Attend to fine details.	1.33	0.47	1.10	0.30
25 Stand for long periods of time.	1.33	0.47	1.40	0.49
26 Work continuously for long periods of time.	1.36	0.61	1.43	0.49
27 Work in a repetitive job.	1.36	0.48	1.46	0.50
28 Work quickly with high accuracy.	1.36	0.48	1.53 A	0.50
29 Work in a confined area for long periods.	1.38	0.49	1.43	0.49
30 Analyze problems quickly and take proper action under stress.	1.42	0.49	1.31	0.46
31 Understand a system of procedures.	1.43	0.62	1.07	0.26
32 Classify objects according to established rules.	1.43	0.49	1.73 A	0.44
33 Apply principles to solve practical problems.	1.50	0.50	1.27	0.45
34 Work in isolation from coworkers	1.50	0.63	1.64 A	0.48
35 Keep work area neat and well-organized.	1.53	0.67	1.53 A	0.50
36 Maintain concentration over long periods.	1.54	0.63	1.46	0.50
37 Use proper spelling, punctuation, and grammar.	1.71	0.70	1.00	0.00
OTHER CHARACTERISTICS...				
38 Honesty.	1.00	0.00	1.08	0.27
39 Emotional stability.	1.07	0.26	1.00	0.00

SD : Standard Deviation. A : Developed primarily after employment as TC-I.

* 5-point scale of Importance, where 1 : Very High and 5 : Very Low.

** 2-point scale of When Developed, where 1 : Before and 2 : After Employment as TC-I.

Exhibit 2. (PART OF APPLICATION BLANK FOR ENTRY-LEVEL TOLL COLLECTOR)

Fill in the "Yes" or "No" circle for each question. If you fail to answer any question, your chances of being hired will be less, and you may be disqualified from further consideration for employment.

SECTION I.

A. Minimum Qualification

Yes No

1 Have you successfully completed the 8th Grade? .... Ⓨ Ⓝ

B. Ability

2 Are you able to count money and make change quickly and accurately? .... Ⓨ Ⓝ

3 Are you able to handle, sort and bundle dollar bills and coins quickly and easily? .... Ⓨ Ⓝ

4 Are you able to reach out with your left hand to collect tolls while pushing a button with your right hand? .... Ⓨ Ⓝ

5 Are you able to read road maps and give proper directions to drivers who are lost? .... Ⓨ Ⓝ

6 Are you able to maintain a high level of concentration and accuracy every minute that you work? .... Ⓨ Ⓝ

7 Are you able to work standing on your feet for up to 7.5 hours per shift? .... Ⓨ Ⓝ

8 Do you have at least 20/30 vision, with correction? .... Ⓨ Ⓝ

C. Public Contact

9 Are you willing to perform work that subjects you to frequent verbal abuse (insults, profanity, obscene gestures, etc.) from drivers who protest having to pay tolls or having to be delayed in traffic backups? .... Ⓨ Ⓝ

10 Are you willing to take verbal abuse from drivers without ever saying anything back or getting into arguments? .... Ⓨ Ⓝ

11 Are you willing to pick up money that drivers have intentionally thrown on the ground in protest? .... Ⓨ Ⓝ

12 Are you willing to collect sticky, smelly, or dirty money that drivers have intentionally "doctored" in protest? .... Ⓨ Ⓝ

13 Are you willing to perform work that will occasionally expose you to flim flam artists, people who give counterfeit money, people who run through the toll booth, and other persons who try to avoid paying their proper toll? .... Ⓨ Ⓝ

14 Are you willing to perform work that will occasionally expose you to people who are diseased or physically deformed? .... Ⓨ Ⓝ

15 Are you willing to perform work that will occasionally expose you to naked people, flashers, and other "weirdos"? .... Ⓨ Ⓝ

16 Are you willing to smile and be pleasant to every driver that comes to your toll booth (up to 4,500 per shift)? .... Ⓨ Ⓝ

17 Are you willing to say "Thank you!" to every driver that pays a toll at your toll booth (up to 4,500 per shift)? .... Ⓨ Ⓝ

18 Are you willing to provide the same high level of service to all persons, without discrimination (by race, sex, etc.)? .... Ⓨ Ⓝ

19 Are you willing to work in a highly visible job in which you will be watched closely not only by your supervisors but also by elected officials, news media, drivers, and the general public? .... Ⓨ Ⓝ

20 Are you willing to accept the responsibility of service as an official representative of the State of Maryland? .... Ⓨ Ⓝ

D. Policy

21 Are you willing to wear a MdTA uniform to work? .... Ⓨ Ⓝ

22 Are you willing to present a neat and clean appearance at the start of each shift? .... Ⓨ Ⓝ

23 Are you willing to work in a job that has strict rules which you must follow at all times? .... Ⓨ Ⓝ

24 Are you willing to perform work that will be closely monitored by your supervisors? .... Ⓨ Ⓝ

25 Are you willing to work for an organization that will closely monitor your use of sick leave? .... Ⓨ Ⓝ

26 Are you willing to pay out of your own pocket to cover any shortages in your cash drawer (due to wrong toll collected, counterfeit money collected, missing money, etc.)? .... Ⓨ Ⓝ

27 Are you willing to work in a job that has a six month probation period, with a possible extension to twelve months? .... Ⓨ Ⓝ

28 Are you willing to work in a job that has limited opportunities for promotion to higher level jobs? .... Ⓨ Ⓝ

DO NOT WRITE IN THIS AREA

E. Availability

Yes No

29 Are you willing to work on weekends? .... Ⓨ Ⓝ

30 Are you willing to work on holidays? .... Ⓨ Ⓝ

31 Are you willing to work the DAY shift (7AM - 3PM)? .... Ⓨ Ⓝ

32 Are you willing to work the EVENING shift (3PM - 11PM)? .... Ⓨ Ⓝ

33 Are you willing to work the NIGHT shift (11PM - 7AM)? .... Ⓨ Ⓝ

34 Are you willing to work a PERMANENT shift (day, evening, or night) as needed? .... Ⓨ Ⓝ

35 Are you willing to work ROTATING shifts (day, evening, and night) as needed? .... Ⓨ Ⓝ

36 Are you willing and able to work up to 16 hours in a row (two shifts back-to-back) as needed in emergencies? .... Ⓨ Ⓝ

37 Are you willing to report back to work on short notice at any time, as needed in emergencies? .... Ⓨ Ⓝ

38 Do you have reliable transportation to work? .... Ⓨ Ⓝ

39 Are you willing and able to come to work on time every day? .... Ⓨ Ⓝ

40 Are you willing and able to come to work in any type of weather? .... Ⓨ Ⓝ

F. Working Conditions

41 Are you willing to perform work that is highly repetitive? .... Ⓨ Ⓝ

42 Are you willing to perform work that demands a high level of concentration and accuracy? .... Ⓨ Ⓝ

43. Are you willing to work in a job that has a lot of pressures to work fast to reduce or eliminate traffic delays? .... Ⓨ Ⓝ

44 Are you willing to perform work that often requires continuous activity, handling up to 500 vehicles per hour? .... Ⓨ Ⓝ

45 Are you willing to perform work that may have long periods of relatively little activity on some shifts? .... Ⓨ Ⓝ

46 Are you willing to perform work that requires standing on your feet for up to 7.5 hours per shift? .... Ⓨ Ⓝ

47 Are you willing to perform work that isolates you from your co-workers most of the time? .... Ⓨ Ⓝ

48 Are you willing to perform work that exposes you to dirt, dust? .... Ⓨ Ⓝ

49 Are you willing to perform work that exposes you to vehicle exhaust fumes? .... Ⓨ Ⓝ

50 Are you willing to perform work that exposes you to loud noises from vehicles? .... Ⓨ Ⓝ

51. Are you willing to perform work that exposes you to some winter cold and summer heat, even inside a toll booth that has both heating and air conditioning? .... Ⓨ Ⓝ

52 Are you willing to perform work that exposes you to some rain and snow, even inside a covered toll booth? .... Ⓨ Ⓝ

53 Are you willing to work in a job that provides only two paid 15-minute breaks and one unpaid 30 minute lunch period during each shift? .... Ⓨ Ⓝ

54 Are you willing to perform work in which you may not leave your toll booth until you have called your supervisor and been given permission to leave? .... Ⓨ Ⓝ

55 Are you willing to perform work handling large sums of money? .... Ⓨ Ⓝ

56 Are you willing to perform work in which someone may point a gun at you or otherwise threaten your life? .... Ⓨ Ⓝ

57 Are you willing to perform work that may expose you to possible traffic hazards in the toll plaza? .... Ⓨ Ⓝ

58 Are you willing to be fingerprinted as part of the background investigation process for job applicants? .... Ⓨ Ⓝ

Provide details in Section IV for ANY question for which you filled in a "No" circle

SECTION II.

Yes No

1 Have you ever been dismissed or forced to resign from a job because you stole something from your employer? .... Ⓨ Ⓝ

2 Have you ever been convicted of a crime in which you stole something that did not belong to you? .... Ⓨ Ⓝ

Provide details in Section IV for EITHER question for which you filled in a "Yes" circle

■○■■○○○○○○○○○○■○■■○○○○○○ 213005

DO NOT WRITE IN THIS SHADED AREA

PEER AND SUPERVISOR EVALUATIONS: AN UNDERUSED PROMOTION METHOD FOR LAW ENFORCEMENT

Douglas Cederblom
Washington State Patrol

In law enforcement, the most commonly used methods for determining promotions to supervisory and mid-management positions are not based on observation of officers' performance on the job. A 1986 survey of 149 law enforcement agencies indicated that to determine promotions to sergeant through captain, 90 percent of the agencies use written examinations, 44 percent use oral examinations, and 32 percent use assessment centers. Only 21 percent use ratings of job performance or promotability (Weiss, 1987). While there are good reasons for using examinations and assessment centers, these methods are nevertheless somewhat contrived, offering a limited view of officers' likely performance at the next level. Peer and supervisor ratings of officers' promotability, based on observation of officers performing in the actual work situation, provide a naturally available and relatively comprehensive method of evaluating promotability.

I would like to discuss several ways in which we have used peer and supervisor ratings to determine promotions at the Washington State Patrol (WSP). We have used various forms of peer and supervisor ratings to evaluate all officers in this agency from cadets through majors. These ratings are then combined with the results of other methods to determine promotions.

Peer ratings and comments were initiated recently to evaluate cadets at the training academy. It was felt that cadets are in an excellent position to observe each other, and that their evaluations of each other would both assist the staff in quickly identifying cadets with problems and help provide feedback to each cadet. Ratings and comments are obtained on three occasions during the twelve-week training. Results indicate enough spread in the ratings to be administratively useful, and the peers' comments are mature and specific enough to be meaningful to fellow cadets. Ratings and comments are included in determining which cadets then become troopers.

For the past ten years, supervisor ratings in the form of forced choice evaluations have been used and weighted significantly (50 percent for troopers, 40 percent for sergeants) to promote these officers to sergeant and lieutenant. A natural concern in using supervisor ratings is possible bias of the rater. With forced choice evaluations, the weights of the rating scales are hidden, so it is difficult for a supervisor to intentionally bias the ratings. To provide a criterion check on the validity of these evaluations, we regularly obtain confidential peer and supervisor ratings of the promotability of the officers, and correlate these ratings with the evaluations and with the other components of the promotion system. Not only do the evaluations correlate highly with confidential ratings by both peers (.50) and supervisors (.57), they also correlate significantly higher with these ratings than our multiple choice written examinations correlate with these same ratings by peers (.27) and by supervisors (.32).

Direct higher-officer ratings of the promotability of sergeants are also used (weighted 5 percent) in determining promotions to lieutenant. Here, the possibility of rater bias is offset by having multiple raters--on the average, five higher officers rate each sergeant. This rating process has been used for the past seven years and has been well-received.

Two years ago, WSP was faced with the likelihood that several higher-ranked individuals would soon retire. Our task was to develop a systematic evaluation of the promotability of lieutenants and above in order to assist the Chief in making needed promotion decisions. Assessment centers had been used twice in the previous ten years to assess mid-level and higher officers. However, this method was costly; participants had complaints about the process; and they did not feel that the results were reflected in subsequent promotion decisions.

It was decided to evaluate lieutenants and above by a combination of peer and higher-officer ratings of promotability. This method would have several advantages: although peer ratings have seldom been used in law enforcement agencies, reviews of their validity in other settings have been very positive (Kane and Lawler, 1978; Lewin and Zwany, 1976); using both sources of raters would provide a large number of raters; and combining different perspectives would provide more complete measures and allow one source to offset possible bias from another source.

The promotability of 60 lieutenants, captains, and majors was appraised in two evaluation processes. In the first process, participants were advised to rate each other's promotability on a 1-5 scale on each of four managerial dimensions (judgment, administrative skills, personal impact, and work involvement). Officers were instructed to rate only those individuals with whom they had frequent job contacts. Each officer received an averaged rating in each dimension, separately from peers and from higher officers, and the officer's relative position among all participants (top, middle, or bottom third), based on the ratings.
In the second process, participants were asked to assume that the positions of the six majors and four deputy chiefs had just become vacant and to nominate the

officers (including incumbents) they thought would do the best jobs in these generic positions. Participants were advised that the results of both processes would be seriously considered by the Chief in making future promotions and assignments.

Results showed sufficient variance for making administrative decisions. Individuals' overall ratings ranged from a high of 4.1 to a low of 2.4, M = 3.3, SD = 0.4. The evaluation processes were reliable and valid. Inter-rater reliability on each of the four dimensions ranged from .90 to .93, with the average lieutenant/captain rated by 29 individuals. Ratings by peers and by higher officers generally agreed (r = .78, and insignificant mean differences between the two groups of raters). High validity was indicated by a significant correlation between the two processes (r = .63), and by a lack of seniority bias. Very few criticisms were heard from officers when they received their individual results. Almost two years later, ten lieutenants and captains have been promoted from the top third, four promoted from the middle third, and zero from the bottom third.

Thus, a reliable, valid process was developed quickly at low cost and used with high acceptance. Yet this kind of process is not often used. Briefly, what are likely sources of resistance to promotability ratings, and facts/ideas for dealing with resistance?

1. Concern that peer ratings are merely a popularity contest. [However, several studies have shown that factoring "friendship" out of peer ratings does not significantly reduce the solid validity of peer ratings (Hollander, 1956; Love, 1981). Apparently some of the same qualities involved in friendship - e.g., people skills, personal credibility - are also important for being an effective employee.]
2. Concern about a low number of raters per ratee. [Combining peers and higher officers (and possibly subordinates) provides a very credible, reliable number of raters.]
3. Concern that the organizational culture is too distrusting to support this method. [To counter this, try promotability ratings with a relatively low weight and combine the ratings with other promotional methods. If the ratings prove effective, increase their weight for subsequent promotion cycles.]

REFERENCES

Hollander, E.P. (1956). The friendship factor in peer nominations. Personnel Psychology, 9, 435-447.

Kane, J.S. and Lawler, E.E. (1978). Methods of peer assessment. Psychological Bulletin, 85, 555-586.

Lewin, A.Y., and Zwany, A. (1976). Peer nominations: a model, literature, critique, and a paradigm for research. Personnel Psychology, 29, 423-447.

Love, K.G. (1981). Accurate evaluation of police officer performance through the judgment of fellow officers: Fact of fiction? Journal of Police Science and Administration, 9, 143-148.

Weiss, J.G. (1987). Statistical summary of responses of 149 police departments to 1986 survey of examination practices for promotion to the ranks of police sergeant, lieutenant, and captain. Unpublished manuscript, Metropolitan Police Department, Government of the District of Columbia.

ISSUES AND PROBLEMS IN THE DEVELOPMENT AND IMPLEMENTATION OF JOB SIMULATIONS

BACKGROUND

The City of New Orleans has used work sample tests for major examinations since 1984. In the past five years, many techniques have been employed to make examinations more job related and to lessen adverse impact. We are sharing with IPMAAC participants some of our success stories and the problems encountered in hopes of improving the examination process.

Over the past five years, we have had the opportunity to develop and administer three major examinations involving job simulation exams. The examinations were for the following positions: Police Sergeant, Police Lieutenant, and Fire Captain. We also developed and administered a job related Police Recruit examination.

In the process of testing, we had several naturally occurring events which allow us to make some observations. For example: (1) Many of the same candidates took both the Police Sergeant and Police Lieutenant examinations, (2) The training sessions differed in degree of complexity. Materials used in training are discussed. (3) Different testing modalities were used to test the same dimensions. These modalities include in-baskets, case studies, role plays, and video observation. (4) Two different rating formats were used. One format was a behavioral checklist. The second was a Behaviorally Anchored Rating Scales (BARS) format. We investigate and make some conclusions about the usefulness of these formats.

THE EFFECT OF HAVING PARTICIPATED IN A WORK SAMPLE ON PERFORMANCE IN A SECOND JOB SIMULATION

In 1984, an examination for the position of Police Sergeant was administered by the New Orleans Civil Service Department. Subsequently in 1988, an examination for the position of Police Lieutenant was administered. Accordingly, many candidates who participated in the Sergeant's examination were eligible to take the Police Lieutenant's examination. Many of the Police Lieutenant candidates were concerned that those who had taken the 1984 Sergeant's examination would have an advantage over the candidates who had not had the opportunity to go through a job simulation. Our staff tried to equalize the effects of job simulation experience by familiarizing all candidates with this type of test. Furthermore, our staff wanted to test the hypothesis of whether prior participation in a job simulation affects future assessment process performance. We used data on all candidates who participated in the Lieutenant's examination to determine if candidates who had participated in the 1984 Sergeant's job simulation exam performed significantly better than those who had not.

THE IMPORTANCE OF TRAINING SESSIONS FOR JOB SIMULATION EXERCISES

Because the use of assessment centers is increasing for job selection, placement, and promotion, whether the testing agency should offer a training session for applicants has become a current issue. Many applicants have been exposed to assessment centers in previous testing situations, but other applicants who have not been tested in several years may never have experienced a job simulation. When these applicants are competing for the same job, mandatory training assures that all candidates are exposed to the same information regarding the test.

Depending on the candidate's training needs, the orientation session can range from a simple information session to an elaborate training session. Other factors which may affect the type of training session administered include cost, personnel available, time, and special conditions such as candidates' test experience.

Training sessions accomplish several objectives. First, the applicant's anxiety may be relieved by supplying test taking strategies and helpful hints on preparing for the test. Second, the candidate will be educated on the dimensions measured by the test. Third, the candidate can become acquainted with the type of job simulations that are involved in the testing process. Sample tests may be given to applicants to familiarize them with the test material. Also, open discussion is usually encouraged which enables the candidates to address any issue they have regarding the entire testing process.

EFFECTS OF DIFFERENT TESTING MODALITIES ON JOB SIMULATION RATINGS OF MINORITY AND NON-MINORITY CANDIDATES

One touted advantage of job simulation exercises for personnel selection is that they provide decreased adverse impact when compared with written examinations. Accordingly, many jurisdictions have turned to work sample examinations for personnel selection to decrease adverse impact. However, there may also be differences between work sample testing modalities in their usefulness for decreasing the disparity in scores between minority and non-minority candidates.

The Uniform Guidelines mandate the use of the selection procedure that demonstrates the least adverse impact against protected classes when compared with other equally valid procedures. Thus, the determination of which types of job simulation exercises show the least difference in scores between minority and non-minority candidates is important to the area of personnel selection.

The purpose of the present research was to explore the effects of different job simulation modalities on the scores of minority versus non-minority candidates for a Police Lieutenant work sample test. Effects of four work sample test modalities were compared within seven job dimensions using seven multivariate analysis of variance procedures. These analyses tested for differences between black and white applicants within each of the seven job dimensions.

Results indicated that there were no significant differences between the scores of black and white applicants for the seven analyses. While more research in this area is warranted, this exploratory study indicated that test modality in job simulations may be less important than it is often presumed to be for reducing score differences between minority and non-minority candidates.

COMPARING BEHAVIOR-BASED RATING FORMATS USED FOR ASSESSMENT CENTERS AND OTHER JOB SIMULATIONS

Despite admonishments that rating format has little effect on the quality of rating (e.g., Landy and Farr, 1980), untold hours are spent in developing elaborate rating forms for job simulations. Rating formats most often chosen are behaviorally anchored rating scales (BARS) and behavioral checklists.

Among the reasons these formats are chosen are that first, they allow those who develop the simulation to anticipate the range of candidate responses and the acceptability of those responses. Thus, the development of behavioral rating forms can serve as a final "review" of a job simulation exercise for test developers. Second, behavioral scales aid in familiarizing assessors with acceptable responses for the position being assessed.

This presentation includes a discussion of the relative advantages and disadvantages of BARS and behavioral checklists for job simulations. Alternative rating formats are also discussed. As it is not likely that format differences would affect the psychometric properties of the ratings, conclusions are based on such factors as ease of use and flexibility.

Anne Russo
City of New Orleans

OBSERVATIONS ON BEING AN EXPERT WITNESS

Nancy E. Abrams

Each of us, except for Mike, have been experts in cases involving Title VII or related employment litigation related to testing. From what I have read, experts in many other areas have similar observations. The first thing you may ask is what do expert witnesses do. They do many other things beside testify.

Preparation

When approached by an attorney involved in a case, I usually do not commit myself. I agree to review all records and orally explain my views. Lawyers rarely want experts to write much down - it could be discoverable by the other side.

To back up, where does business come from. Strictly expert witness work comes from attorneys. They usually get my name from other attorneys I have worked with. Often attorneys, especially defense attorneys, have limited knowledge of this field of expertise.

Attorneys want you to support their position 100% - usually this is impossible. You must discuss your feelings of strengths and weaknesses of the case. It is essential that they understand what you are willing to say or not say before you are formally retained in the case. You often can suggest additional studies or analysis which may help the case. Once retained, you work on these.

You should work with the attorney on strategy, other sides experts, etc. It helps to understand some of the legal jargon. You also must be able to explain our technical terms in non-technical ways. You must always be able to communicate with the attorney. If this breaks down or the attorney starts playing expert you may be in for trouble.

Depositions

In some cases, an expert may be deposed. This means that the attorneys from the other side ask you questions under oath. They use this to better understand the opposing side's case. You may also assist your attorney with experts from the opposing side, developing questions and explaining answers.

Settlement Negotiations

Often experts get involved in settlement negotiations. You may be asked to evaluate the technical aspects of proposed settlements. You must explain the impact of such proposals to attorneys and their clients. Sometimes you may be asked to attend negotiating sessions.

Irial Iestimony

If you and your attorney are well prepared, your direct examination should go smoothly. You should know just what to expect. Cross examination is often unexpected. Your job is to try to plan for what to expect. An important idea to remember is that attorneys try not to ask questions unless they know the answer. Sometimes cross examination is very short, other times long and labored. Also it may be friendly or it may be very hostile.

It is important to remember that judges often ask questions. These are usually friendly but are often critical. They tell you where the judge is coming from and what he understands or doesn't.

You also may assist in the cross examination of experts from the other side by again preparing questions for the attorney and explaining answers.

Observations

Being an expert witness can present ethical concerns. Attorneys sometimes try to make you say things they want you to say. You must educate the attorneys on why you must say what you say.

Being an expert witness can be rewarding but also frustrating. The better prepared you are, the better you will fare.

LET'S GIVE LAWYERS THE RESPECT THEY DESERVE

Lance W. Seberhagen, Ph.D.
Seberhagen & Associates*

Lawyers have been held in low regard for centuries. Here is a sampling of what people say:

1. "A lawyer is someone who is skilled in the circumvention of the law" (Roth & Roth, 1988).

2. "A lawyer is someone who helps you get what's coming to him/her" (Roth & Roth, 1988).

3. "A lawyer is a person who profits by your experience" (Roth & Roth, 1988).

4. "A lawyer is a person who writes a 10,000 word document and calls it a brief" (Roth & Roth, 1988).

5. "Lawyers sometimes tell the truth--they will do anything to win a case" (Roth & Roth, 1988).

6. "Law is the exact opposite of sex. Even when it's good, it's lousy" (Roth & Roth, 1988).

7. "No lawyer will ever go to heaven as long as there is room for more in Hell" (Roth & Roth, 1988).

8. "What is the definition of a tragedy? A bus full of lawyers going over a cliff with two empty seats" (Landy, 1989).

9. A lawyer was giving a talk to a group of law students about what it's like to practice law. One student asked the lawyer if he ever faced any serious ethical problems in his practice.

   LAWYER: Let me give you a specific example. Once a client was billed for $5000 worth of legal work but sent me a check for $8,000.

   STUDENT: What's the ethical problem?

   LAWYER: Whether to tell my partner! (Schwartz, 1989)

10. Why are they now using lawyers instead of rats for laboratory research?
    a. There are more lawyers than rats.
    b. There are some things that rats just won't do.
    c. Some people develop an emotional attachment to rats. (Anonymous)

I have worked with many lawyers in the course of my work. Lawyers are not all bad, but I do have a few complaints (see Table 1) that may give some insights about what it is like to be an expert witness.

* 9021 Trailridge Court, Vienna, VA 22182. Tel. (703) 790-0796.

Table 1

THE TOP TEN COMPLAINTS BY EXPERT WITNESSES ABOUT LAWYERS

SOME LAWYERS...

1. Don't pay expert's fees for many months or years, if at all. (NOTE: Applies primarily to plaintiff's lawyers.)

2. Ask surprise questions to expert on the witness stand.

3. Don't object to dirty tricks to expert by opposing counsel during cross-examination.

4. a. Re-phrase questions written by expert, missing point of original questions.

   b. Don't ask all questions written by expert, omitting critical set-up questions or ultimate target questions.

   c. Don't introduce all exhibits prepared by expert.

5. Won't ask question to witness unless they already know answer, even if helpful answer is likely and worst possible answer wouldn't hurt case.

6. a. Don't tell expert about plans for presenting case.

   b. Don't tell expert truth about case.

   c. Never give expert straight answer to any question.

7. Won't let expert write report for presentation to court, forcing expert to present detailed facts and complicated analyses in oral testimony.

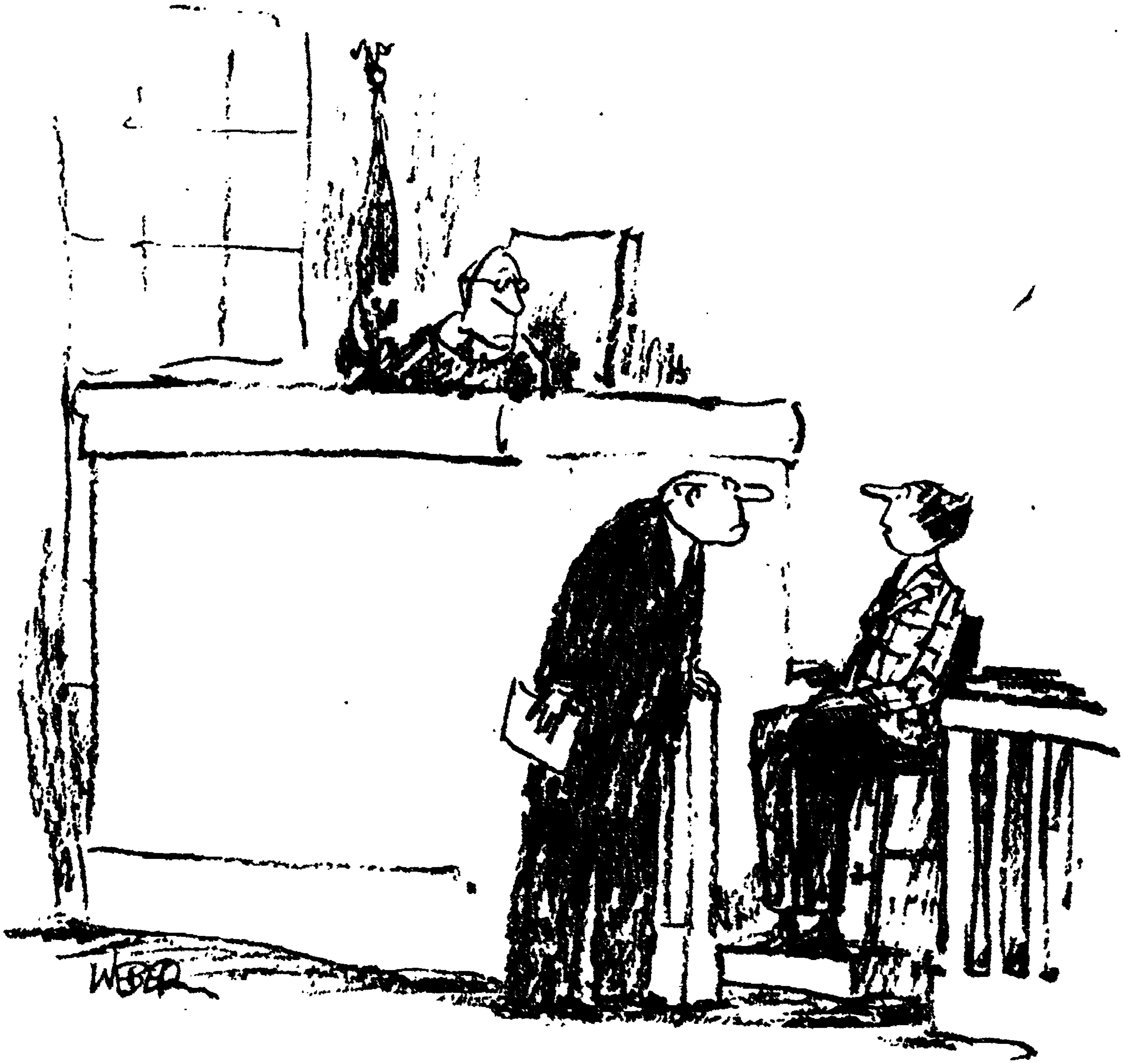
8. Depend on expert to interpret law.

9. a. Call expert at last minute, giving expert little time to prepare for trial.

   b. Call expert to work on case after depositions have been taken from witnesses or deadline has passed for requesting case documents.

10. Name expert for case before contacting expert, or reaching agreement with expert, getting unauthorized use of expert's name and reputation.

NOTE: Listed in rank order, where #1 is biggest complaint.



"Your Honor, I feel threatened by this gentleman's intensity."

Agency-Rated Banded Scores: An Alternative to Traditional Educatior and Experience Ratings for Highly Specialized Job Classes

Presented at the Thirteenth Annual Conference of the International Personnel Management Association Assessment Council at Orlando, Florida, June 1989

8:30 a.m., June 20, 1989

Ann R. Williamson, Senior Exam Development Analyst
Ronald A. Ulm, Project Leader
State of Georgia Merit System
200 Piedmont Avenue, Suite 418
Atlanta, Georgia 30034

INTRODUCTION

One of the traditional advantages of a selection procedure which involves a rating of education and experience is the expedience with which it may be developed and administered. However, in order to make assessments, the applicant evaluator must be able to understand the rating criteria, as well as the information the applicant provides. If the job class at hand is highly specialized and requires specific technical knowledge, the assessment becomes more difficult to make. In such cases, applications will often abound with technical jargon typically unfamiliar to one outside of the field. Furthermore, these highiy specialized classes often generate the least applicant activity, thereby reducing the cost-eftectiveness of extensive rater training.

The State of Georgia Merit System determined that, for some highly specialized classes, the staff of the hiring agency itself might prove to be the best evaluator of applicants. The Merit System set out to develop a content valid instrument which would allow Subject Matter Experts (SMEs) from the hiring agency to assess the applicants; grcup them into broad ranking categories; and conduct the project in less time than normally required for development and implementation of a traditional rating of education and experience. The procedure, known as the Agency-Rated Banded Scores Instrument (ARBSI), has been in place since March, 1988, and has been utilized by five job classes, thus far.

The ARBSI appears to be most appropriate for job classes which satisfy the following criteria:

1. Due to agency time involved, the class must receive few qualified applicants. In addition, the ARBSI is vacancy-specific; therefore classes with low hiring activity are the most suitable.

2. In order to justify the departure from traditional applicant evaluators, the class should be highly specialized, with a well-developed professional jargon.

3. The project must be a high priority for the hiring agency. The ARBSI may be developed and administered in a relatively short time, but only with full agency cooperation.

As an example, the Merit System recently utilized the ARBSI to assess applicants for four vacancies in the class Systems Analyst, which were announced to the public last year. For these particular vacancies, extensive additional recruitment had taken place to the extent that seventeen applicants appeared to qualify. Systems Analyst made a particularly good candidate for the ARBSI because it easily met all of the established criteria. No selection procedure existed with which the applicants could be evaluated.

METHODOLOGY

First, the agency personnel staff was contacted. The Merit System explained the ARBSI and the necessary level of agency commitment. Then, the agency personnel staff was instructed to provide the three supervisors of the vacant positions with class specifications and the four position descriptions. The supervisors, who served as departmental SMEs, were instructed to review these materials and identify knowledge, skills, and abilities (KSAs) necessary for performing the duties of the Systems Analyst vacancies they supervised.

Next, a meeting was held with Merit System Exam Development Analysts, the Subject Matter Experts (SMEs), and the agency personnel staff. Relevant background information on the project was provided to the SMEs and they were briefed on their roles in evaluating the applicants. Then, the SMEs were asked to discuss the knowledge, skills, and abilities (KSAs) they had established as necessary to perform the duties of the Systems Analyst vacancies they supervised, based on the job materials they reviewed.

The information gathered from the SMEs was used to construct a three-part form by which the seventeen applicants would be evaluated. Section I of the form was a detailed adaptation of the minimum qualifications for the class. If the applicant did not meet the requirements of this section, then the evaluation would end. Applicants who appeared to qualify would go on to be evaluated in the remaining two sections.

Section II, the KSA Rating Section, asked the rater to assess the extent to which the applicant possessed each of the previously-defined KSAs. The four-point rating scale was: 0=none, 1=minimal, 2=average, 3=above average. Since many of the KSAs generated were quite specific to Systems Analyst duties, no avenue existed for crediting more general electronic data processing backgrounds. Therefore, the final section asked the rater to make an overall assessment of the job-relevance of the applicant's total education and experience background. The SMEs would be required to document any additional information, other than extent of KSAs possessed, used in making this determination (such as previous experience with hardware/software used by the state, familiarity with specific systems, etc.). The overall qualifications scale was benchmarked as: Unqualified, Minimal Qualifications, Average Qualifications, and Excellent Qualifications.

After the rating form was finalized, a second meeting was arranged with the SMEs. During this second meeting, the raters were given the applications of the seventeen potentially qualified candidates, along with the newly-developed rating forms attached to each. The SMEs were trained in proper utilization of the evaluation instrument and in the method by which their ratings would correspond to the applicants' final score bands. Then, the SMEs were informed that the top two score bands -- both the "Excellent" and "Average" groups would compose the list of eligible candidates for selection.

RESULTS

The completed rating packages were returned by the Subject Matter Experts within two days. A Quattro spreadsheet was used to analyze rater data. For each rater, the applicant's name, rating for each KSA, the mean for all KSA ratings, and the overall rating were entered. Then, for each applicant, the overall rating and the mean KSA ratings were averaged across all three raters. These two means were averaged to produce the applicant's final score band.

To determine the relationship between the KSA-based scale and the overall qualifications scale, the Pearson Product Moment correlation coefficients were calculated for each rater. Intra-rater reliability for the two measures was found to be highly positive (r=.88, r=.83, r=.86). These strong relationships indicate that the KSA-based scale and the overall qualifications scale are measuring highly similar dimensions of the relative worth of the applicants' backgrounds.

The final scores had a range from 2.0 to 2.9. It was somewhat unusual that seventeen applicants received final scores of 2.0 or better on a scale of 0 to 3. Since the SMEs were informed in advance that applicants with scores of 2.0 or greater would be eligible for consideration, they may have been operating on a true scale of 2.0 to 3.0, thereby affording consideration to all applicants[1]. Those applicants who scored 2.6 or higher were placed in the "Excellent Qualifications" group; those with scores from 2.0 to 2.5 were placed in the "Average Qualifications" score group.

DISCUSSION

In conclusion, the ARBSI has been a valuable addition to the instruments available to the Merit System for applicant evaluation. There are, however, some restrictions. Considering the amount of agency time involved in performing the rating, it would probably be wise to limit the ARBSI to classes with twenty applications or less. Furthermore, based on the five classes which have used the ARBSI, it appears to work best with technical classes (electronic data processing, engineering, etc.), rather than classes in the social service areas. The inferences made by SMEs in evaluating applicants for the latter classes appeared to be more subjective; the SMEs "reading into" the applications qualifications not clearly apparent from the description of duties.

Overall, this instrument provides many advantages not available in a traditional rating of education and experience. First, it took a remarkably short amount of time to develop, utilize and implement: approximately three weeks from the first SME meeting to release of a list of eligible candidates. In addition, the instrument allowed a finer assessment of technical qualifications than that which a rater would normally make. The instrument also allowed distinctions to be made in the quality of the applicants' experience, rather than just duration.

Nevertheless, some disadvantages to the ARBSI remain. First, the instrument has a rather minimal linkage to tasks in establishing validity. Further, it is difficult to control for rater bias, both individually and as a group. In future research efforts, strong consideration should be given to attenuating for rater self-interest. Perhaps a number of applications could be fabricated to resemble applicants with excellent, average, poor, and even no qualifications. The ratings assigned to the decoy applications might indicate the true scale under which the rater was operating. Then, the final scores of the real applicants could be adjusted accordingly.

[1] While the Systems Analyst results appear skewed toward the high-end of the scale, the other classes which used the ARBSI had more normal distributions.

Development Of Valid Computer-Based Tests For Assessing Divided Attention In Personnel Hired As Operators Of Nuclear Power Plants

The presentation describes a recent development in the area of computer-based personnel testing. A new test, which is compatible with IBM-PC's, has been designed to assess divided attention skills and work load capacity in people who apply for jobs as operators in nuclear power plants. Applicants who do well in this test can successfully perform several critical tasks at the same time and can maintain this level of effective performance when facing the "stress" of increases in the work load during subsequent stages in the testing. These are critical skills which contribute to a person to being able to perform the nuclear operator's job tasks in an effective and safe manner.

The test battery in the validation involves the Divided Attention test and several traditional, paper-pencil tests (e.g. PTI Numerical, Bennett Mechanical, and FIT Assembly). Inclusion of these tests in the battery is based on the results of an extensive job analysis which establishes the skills and abilities (i.e. divided attention, spatial, mechanical comprehension, numerical, and clerical perception) personnel must have in order to operate in nuclear power plants effectively.

The Divided Attention test assesses the person's ability to perform several tasks at the same time. These include keeping a pressure indicator within "safe limits" (i.e. rate control), responding quickly and accurately to different tones (i.e. reaction time), remembering different operating procedures with their appropriate c·des, and estimating the amount of time elapsing during different stages in the testing process. In taking the test, each applicant first performs both the tracking and the reaction time tasks separately for three trials and then performs these tasks at the same time during the remaining eight trials.

A concurrent, criterion-oriented method was used in the validation. The sample included 248 incumbent operators. First and second level supervisors completed a performance appraisal form for each incumbent. They rated the operator's job performance in terms of 22 dimensions found to be important aspects of job effectiveness during the job analysis.

The results indicate that all of the tests in the battery are valid. For the Divided Attention test, the test - retest reliabilities are high enough to indicate implementation is appropriate. Although the Divided Attention test does not have as high of validity as several other tests in the battery, it emerges as the second best predictor because the Divided Attention test is relatively independent with the other tests in the battery.

David C. Myers
Manager, Assessment
Services, Duke Power Co.

and

Mark Schemmer
University Research
Corp., Bethesda, MD

Strategies For Making Cut-Score Determinations On A Performance-based, Observational Test

I. Leon Smith and Sandra Greenberg
Professional Examination Service

From a standard setting or cut-score perspective, performance-based tests represent a unique challenge because the instruments are quite different in conception and scoring from the multiple-choice format for which most cut-score methods have been developed. This paper focuses on a description of a cut-score methodology designed for use with a classroom observation instrument for beginning teachers.

The items (called *indicators*) in the classroom observation instrument differ in two important respects from the multiple-choice question format. First, with regard to multiple-choice questions, getting an item correct represents a *score*, but does not imply a standard of performance. On the classroom observation instrument, getting an acceptable or unacceptable rating on the behaviors associated with each of the ten (10) indicators is both a score *and* a standard of performance. That is, a trained assessor must make a professional judgment as to whether the teacher has sufficiently demonstrated the indicator-associated behaviors to earn an acceptable rating.

Second, the items selected for placement on multiple-choice examinations are typically intended to randomly or representatively sample particular domains or categories of performance. The ten indicators on the classroom observation instrument, however, are not conceptualized as samples of some larger domain but as a small, but comprehensive, set of critical dimensions of teaching based on the collection of rational (subject-matter expert opinion) and empirical (job analysis) data.

Cut-Score Strategies

The standard setting procedures are based on the assumption that each teacher will be observed six (6) times on the ten (10) indicators and rated *Acceptable* or *Unacceptable* on each indicator.[1] It is also assumed that the observations will span different topics and be conducted by different observers. The observations in the Fall (first three sessions) and Spring (second three sessions) will be spread out rather than bunched together within short periods or rounds. The goal of standard setting is to come up with a formula for combining the 60 separate *Accept/Unaccept* ratings into a single pass/fail decision.

There are two general approaches to such a formula. The first approach is referred to as *observation-focused* and would require the beginning teacher to demonstrate competence during a proportion of the six observations. The second approach is referred to as *indicator-focused* and would require the beginning teacher to demonstrate an acceptable level of competence with regard to the ten indicators.

Although both approaches have merit, it would appear that the *indicator-focused* method is most consistent with the purposes and goals of the teacher observation instrument. The indicators are not random samples from some larger performance domain, but represent the essence of teacher competence as defined through rational and empirical analysis. An

[1] In the case of one (1) indicator, called LESSON CONTENT, a *Can Not Rate* rating may be appropriate.

observation-focused approach would appear to divert attention away from the primary emphasis in the teacher behavior instrument. Accordingly, a strategy was designed to produce two slightly different formulas based on the *indicator-focused* method. The first formula is *non-compensatory*, and would require that the beginning teacher demonstrate an acceptable level of competence on *each* indicator in order to pass. The second formula is *partially compensatory* and would require the beginning teacher to demonstrate competence on essential indicators *and* on a percentage of the remaining indicators.

Meeting 1 of the Standard Setting Committee

Ratings data related to the two formulas for setting standards will be collected from a panel of subject-matter experts familiar with the teacher observation instrument.

Each panelist will be required to answer the following seven (7) questions. Special rating forms will be prepared to permit the panelists to make choices from a set of predetermined options.

Q1. *For each of the ten (10) indicators*, circle the number of *Acceptable* ratings a beginning teacher must obtain in order to be judged competent on that indicator. The choices are 6, 5, 4, 3, 2, or 1. Your answers may vary for any and all of the 10 indicators. The panelists will be reminded that one indicator can be scored as *Can Not Rate* and that this should be considered when making their ratings on that one indicator.

The second question requires the panel members to consider the *timing* of the ratings.

Q2. For a beginning teacher to be judged competent on each of the indicators, circle the *minimum* number of *Acceptable* ratings that the teacher can earn during the *second three (3) observation sessions*. To answer this question, you will have to refer back to your answers in Q1. The choices are 0, 1, 2, and 3. Your answers may vary for any and all of the indicators. For example: If you answered that a teacher had to earn 6 *Acceptable* ratings on any indicator over the course of all six (6) observations, it is clear the teacher will have to earn 3 *Acceptable* ratings during the *second three observation sessions*. If you answered that a teacher had to earn 5 *Acceptable* ratings on an indicator, then you must select a minimum of either 2 or 3 *Acceptable* ratings during the *second three observation sessions*. If you answered that a teacher had to earn 4 *Acceptable* ratings, then you must select either a minimum of 1, 2, or 3 *Acceptable* ratings during the *second three observation sessions*.

The third question deals with the one indicator for which a *Can Not Rate* may be appropriate.

Q3. For the LESSON CONTENT indicator, circle the *maximum* number of *Can Not Rates* beyond which a rational decision concerning teacher performance should *not* be made. The choices are 6, 5, 4, 3, 2, or 1. The panelists will be reminded about the reasons for the *Can Not Rate* category and the experiences utilizing the rating during the pilot administration will be discussed.

The next question requires the panelists to consider the impact of each indicator on the overall pass/fail decision.

Q4. Consider each indicator, one at a time, and indicate if you believe a beginning teacher *must* be competent on the indicator in order to pass the classroom observation instrument. Circle *Yes* if you believe that the teacher must be competent on the indicator in order to pass, regardless of that teacher's performance *on any and all of the other indicators*. Circle *No* if you do not believe that the teacher *must* be competent on the indicator in order to pass.

Question five requires the panelists to rate the indicators they believe are not essential to passing the teacher behavior instrument.

Q5. If in Q4, you said *Yes* to all of the indicators, skip this question. If not, review your answers to Q4. Assume that a beginning teacher was competent on all of the indicators you rated as essential to pass. Circle the number to indicate how many of the *remaining* indicators the teacher must be rated as competent in order to pass the teacher behavior instrument. Conversely, on how many of the remaining indicators can the beginning teacher be rated not competent, and still *pass*? Circle the number to indicate how many of the *remaining* indicators the teacher can be rated as not competent, and still *pass*. Your choice for both of these questions is any number between 0 and 10, but the *combined* total of your answers must be equal to the number of *No* answers in Q4.

The panelists will then be asked to make a global estimate of the passing rate on the teacher behavior instrument.

Q6. Circle the percentage of all beginning teachers in the state you believe will successfully complete this phase of the teacher certification program.

After collecting all of the ratings data based on the six (6) questions, the results will be tabulated. Special summary sheets will be developed for this purpose.

The panelists will then determine whether to retain or revise their original standard setting ratings based on discussions with the other members of the panel. The initial ratings' data will be summarized along with information from a pilot administration[2] and a job analysis study.[3] With these data in hand, the panel members will decide whether their initial ratings merit revision.

Finally, panel members will be polled to identify the standard setting formula they believe will make the most sense to the relevant constituencies. Specifically, the panelists will be asked:

Q7. Which formula, the *non-compensatory* or the *partially compensatory*, do you feel is more meaningful in assessing a teacher's overall competence?

[2] In the pilot study, approximately 100 beginning teachers were observed 6 times. The 100 teachers represented the following peer families: (1) Elementary, (2) Special Education, (3) Vocational Education, (4) Academic (high school), and (5) Special Subjects (Art, Music, Physical Education, and Health). A small group of experienced and experts teachers were also observed for comparison purposes.

[3] A job analysis survey was designed to measure the importance, frequency, and point of acquisition of the indicators comprising the teacher behavior instrument and the underlying teaching competencies. Approximately 1600 beginning teacher completed the survey.

Following the meeting of the standard setting panel, the *non-compensatory* and *partially compensatory* formulas will be applied to the data collected in the pilot study. The impact of both formulas on the pass rate of all teachers observed in the pilot study will be documented.

Meeting 2 of the Standard Setting Committee

The formulas will be reviewed at a second meeting of the panel and three (3) sources of validational information will be presented before deliberation on a final recommendation.

First, three pairs of 6 by 10 matrices will be presented to the panel members. Each matrix will contain either an acceptable (A) or an unacceptable (U) rating on each of the ten indicators for each of the six sessions. One matrix of each pair will be labeled A and the other B.

The total number of acceptable ratings will be the *same* for both matrices in each pair. The pattern of unacceptable and acceptable ratings, however, will *differ*. In one matrix, the candidates will be marginally competent on all of the indicators; in the other matrix, the candidate will be clearly competent on the essential indicators but not competent on some of the non-essential indicators. The panel members will then rate each of the three pairs of matrices in terms of which candidate they believe is most competent.

Two additional patterns of ratings referred to as C1 and C2 will be constructed. The C1 and C2 matrices are *not* paired. The pattern of acceptable and unacceptable ratings in C1 will yield a pass decision based on the partially compensatory approach but a fail decision based on the non-compensatory approach.[4] The total number of acceptable ratings in the C1 matrix will be the *minimum* required to pass at the partially compensatory standard. Matrix C2 will have the minimum number of acceptable ratings required to pass using the *non-compensatory* standard. The panel members will then rate the C1 matrix *and* the C2 matrix in terms of which decision (pass or fail) is most appropriate.

Second, pilot study data documenting the actual number of beginning year teachers who would have been judged competent via each formula will be presented. At the same time, relevant data from Q6 and Q7 collected at Meeting 1 will be reviewed.

Third, pilot study data documenting the actual number of experienced/expert teachers who would have been judged competent via each formula will be presented. Members of the panel will be able to contrast the actual numbers of beginning and experienced/expert teachers who would have passed via either formula as well as the distribution of actual scores.

The final step would be to ask each panelist to select the formula that provides the more reasonable fit to the data. That is, in terms of the validational data described above, the formula that: (a) identifies the more competent candidates illustrated by the three pairs of contrasting matrices A and B; (b) leads to more appropriate decisions based on Matrix C1 and C2; (c) produces overall competency rates closest to the percentages predicted; and (d) is more consistent with realistic expectations regarding the possible differences in performance among beginning and experienced/expert teachers.

[4] It will not be possible to produce patterns that will pass the non-compensatory standard and fail the partially compensatory one.

READING COMPREHENSION AND THE COGNITIVE PSYCHOLOGY OF FRANK SMITH: A GUIDE TO PRACTICE?

IPMAAC National Conference
Orlando, Florida
June, 1989

Michael J. Dollard
New York State Department of Civil Service
Harriman Office Campus
Albany, NY 12239

INTRODUCTION

I was drawn to this topic when I was first assigned to supervise the Verbal Abilities Examining Unit of the New York State Department of Civil Service. As I worked to acquaint myself with the operations of the unit and with the testing materials they used I was struck by what I considered to be anomalies. The test items looked very good, and met all of the basic tests of multiple-choice item development. Further, the item statistics looked very good from a classical test analysis perspective. Yet when the items were combined into our conventional 15-item subtest format, the alpha coefficients were only mediocre: high .60s and low .70s. Later, when I tried to do some further study, I found that inter-item correlations were much lower than I anticipated, barely above .05 levels of significance. This confused and perplexed me. The results were not at all consistent with the little that I knew -- or thought I knew -- about reading. This sent me off to do some reading.

FRANK SMITH

In my search to learn more about reading and reading assessment, I ran into the books of Frank Smith. Smith was born and grew up in England. Following a ten-year stint as a reporter and editor in Europe, Smith's fascination with language brought him to an undergraduate degree from the

University of Western Australia and a Ph.D. in psycholinguistics from Harvard in 1967. For the past twenty years he has been engaged in teaching, research and writing in the areas of language and comprehension, most of it in Canada. Among his large number of books are *Psycholinguistics and Reading* (1973), *Comprehension and Learning* (1975), and *Understanding Reading* (1978).

PSYCHOLINGUISTIC THEORY

It is not my intention to rehearse psycholinquistic theory here. To do so would take weeks if not months, and is beyond my competence. I merely want to sketch out a rough outline of the basic psycholinguistic model, and the implications that Smith draws from it. It is these implications that I believe shed real light on what I am trying to do in assessing the reading skills of candidates for employment in New York State.

The basic psycholinguistic model posits that there are two aspects to language: the *surface structure* and the *deep structure.* The link between the two is grammar (and syntax).

"The sounds of speech and the visual information in print are surface structures of language and do not represent meaning directly. Meaning is part of the deep structure of language and must be contributed by listeners and readers. Reading is not "decoding to sound." Written language and spoken language are related, but are not the same. (*Understanding Reading*, p. 83)

Comprehension of both spoken and written language is rooted in prediction. This process of prediction is the prior elimination of unlikely alternatives. By minimizing uncertainty in advance, prediction relieves the visual system and memory of overloading in reading. Predictions are questions that we ask the world, and comprehension is receiving answers. If we predict appropriately, we comprehend. If we cannot predict, we are confused. If our predictions fail, we are surprised. (*Ibid.*, p.67)

These predictions are based on non-visual information, long-term memory, and prior knowledge: the theory of the world within the individual that is the source of all comprehension. (*Ibid.* p. 67)

Smith uses the term "cognitive structure" to refer to this accumulated totality of all a person's knowledge of the world: this "theory of the world within the head." He argues that this totality of a person's knowledge is not just stored somewhere as a collection of unrelated facts, figures and images, but is stored as a structured or organized collection of information. (He hypothesizes that the difference between poor learners and good learners may not be the *amount* of material remembered, but the degree to which such information is organized and integrated.) (*Ibid.* p. 11)

In general, what Smith is saying is that this cognitive structure, this "theory of the world within the head" is composed of categories (e.g., *dog*) with associated "feature lists" (e.g., *four-legged, furry, medium-size*, etc.) and "links" to other categories (e.g., *isa animal, Rover is an instantiation*, etc.).

It is these categories that I am interested in. Smith sometimes speaks of these categories as "conventions." I think that this is most apt, since it implies that these categories have no necessary "essence", -- as philosophers use that term -- but are in fact learned from experience. In fact your convention *dog* may not be the same as my convention *dog*, particularly if the only dog you have ever experienced is a Bernese Mountain Dog, and my experience is limited to my nine-pound smooth-coat Bruxelles Gryphon.

The term "convention" also permits use with complex pieces of learned behavior. For example the "script" for passive voice is a convention most (but not all) of us learned somewhere around tenth grade that the construction *>was< >verb< >prepositional phrase<* signifies passive voice: the main noun (subject) receives the action of the verb. Another convention many of us learn is that "it's (with an apostrophe)" shows a contracted form rather than a possessive. There are an almost unlimited number of such "conventions": grammatical forms, accepted usages, connotations and denotations of specific terms, idiomatic usages, scripts (learned patterns of words) that convey attitude, perspective or point of view, scripts that signify comparison or contrast, scripts that signify cause or effect, etc.

The key concept is that because possession or non-possession of a specific "convention" is rooted in the individual's specific experience, every individual possesses a different set of conventions; "good" readers have possession of more conventions than poorer readers, but not all "good" readers necessarily possess the same conventions.

THE BROWN UNIVERSITY AND AMERICAN HERITAGE CORPUSES OF AMERICAN ENGLISH

If it is true that reading "power" is based on the possession of these "conventions" which are differentially distributed across the population, how does this fact interact with language that has to be read? For an answer to this I looked to the two modern corpuses of the English (American) language: the Brown University Corpus assembled in the late 1960s, and the American Heritage Corpus assembled in the early 1970s. Both of these corpuses are comprised of large samples of written language that were collected according to a definite sampling plan in order to accurately represent certain classes of written English expression.

Although the two corpuses were collected at different times for different purposes, and despite the fact that neither one of them collected language samples from employment contexts, the findings of the analyses of the two corpuses are similar, and -- I believe -- instructive.

In short, the analysis of both corpuses show that different kinds of reading require different conventions. Although this now seems obvious to me, it was not that obvious to me that long ago!

The American Heritage Corpus (collected as the basis for a new school dictionary, and based on writings used or usable in American schools, public and private, urban and rural, across the country) consists of over 5,000,000 words drawn in 500-word samples from 1,000+ published materials. The first 1,000 types in the token distribution (i.e., roughly the 1,000 most commonly used words) accounted for about 74% of all the words in the corpus; the first 5,000 types accounted for about 89% of the words in the corpus. This study identified 17 different categories of reading materials. Over half (54%) of the 86,000+ types (i.e., different words) were found in only one of the 17 categories of reading materials.

The Brown University Corpus consists of 500 2000-word (average) samples. This study identified nine genres of "Informative Prose" and six genres of "Imaginative Prose." In short, the study found many differences between the two gross categories, and among the various sub-categories.

In looking at sentence length, for example, the Brown University study found mean words per sentence ranging from something over 12 (Mystery/Detective Fiction) to to something over 24 (Miscellaneous Informative Prose {This includes Government documents.}), with an overall mean of something over 18 words per sentence.

Similarly, the Brown University study found that the percent of passive predications ranged from something over 3% (Romance/Love Story) to something over 24% (Miscellaneous Informative Prose); the percent of perfect tense constructions ranged from a low of less than 5% for Skills/Hobbies materials to somewhat over 7.5% for Mystery/Detective materials, and the percent of progressive forms (usually signaling ongoing activities, or as a temporal frame for some other event [e.g., When Paul left, Mary was playing the piano.]) ranged from about 1.5% for Learned materials to almost 4.5% for Press Editorials.

This data is, of course, only a minute sampling from the work done on the two corpuses, but is generally reflective of their findings. It is included here as support for Smith's notion of "conventions" and the need for different conventions (i.e., connotations/denotations of terms, 'scripts' for various constructions/usages, etc.) to comprehend different material.

NEW YORK STATE TEST DATA

As I said at the beginning, the thing that got me interested in this whole issue was my inability to understand the test data I was seeing on the Reading Comprehension materials we were using in the New York State Department of Civil Service. In Appendix 1, I have included item analysis for a typical NYS 15-item reading comprehension subtest, and in Appendix 2 a typical "higher level" reading comprehension test item. While reasonable people may argue about specific aspects of both the test item and the item analysis, I believe that most people in the business would judge the test item to be at least "acceptable" and judge the individual "IAs" as reflecting at least "acceptable" test items. (The statistic IRI is Gulickson's Item Reliability Index and reflects both difficulty and discrimination. The statistic P is the proportion of the candidate population answering the question correctly. The item analysis is based on a median split based on the total score for the subtest, in this case 15 items. The starred frequencies are the intended key for that item.)

My evaluation of the test item and the IA is, of course, more positive than just "acceptable." While I would like to see the items more consistently with IRIs of 0.2 or greater, and Ps of more than 0.5 and less than 0.8, I consider the IA to be "good" and the item to be "well constructed and appropriate." In fact, when I first started looking at this material similar to that in the appendices, I believed we were doing a pretty good job of evaluating the reading skills required by our jobs.

That is why I was concerned with the low KR_{20}s I was seeing in this material. In the IA in the appendix the KR_{20} is .69. This is pretty much a median figure for material of this type, with KR_{20}s normally in the upper 0.60s and lower 0.70s. This is in contrast with most of the general abilities material that my staff develops, where KR_{20}s run from the upper 0.80's to mid 0.90s. In fact, even in a rubric we call Office Record Keeping that typically uses three separate problems with five multiple-choice questions each we typically achieve KR_{20}s in the high 0.80s.

To explore this issue a bit further I used what are essentially *phi* coefficients to estimate inter-item correlations. Recognizing the ceiling effect of unequal item difficulties, I nevertheless expected to see reasonable numbers of coefficients of 0.5 or better, indicating that at least 25% of item variance was being explained by some underlying common factor I was ready to call "reading ability." What I found was that, in subtest after subtest with items and IA like those in the Appendices, typically only about half the inter-item correlations were significantly larger that 0 at the 0.05 level, with the largest correlations running in the mid-0.30s.

Although I wasn't very happy with these results initially, in the light of Frank Smith's theory, and findings of the Brown University and American Heritage studies, these findings are what I would expect to see. If reading ability is in fact "possession of a number of learned conventions about language," and if 'good' readers possess more conventions than 'poor' readers, but if not all 'good' readers necessarily possess the same conventions, then I would certainly expect to see individual test items work well from a classical item analysis point of view (iri>.10, 0.50=<p<=0.80) but with only moderate inter-correlations and alpha coefficients.

IMPLICATIONS OF SMITH'S THEORY FOR READING-ASSESSMENT PRACTICE

The work on the two corpuses seems to me compelling that different types of reading material do require different "conventions" (e.g.,connotations/denotations of different terms, different scripts for different usages, construction, etc.) for adequate comprehension. If it is further true that different persons possess different conventions, then to select persons to do specific jobs which require the reading of specific types of materials, it would seem necessary to map both the conventions required by the types of reading to be done, and the conventions possessed by the potential employees, and then to match the two in some way. At this point I would not be willing to say that every position or even every class of positions differs from every other position or class of positions and would require a separate mapping. In fact the findings of the corpus studies that a relatively small number of word "types" account for a major percentage of the total words in the sample argues against this. But the data would seem to suggest that at the very least the reading skills required of clerks are in fact different than those from paraprofessionals, and that it may well be that the reading skills required of general office clerks are different from those required of account clerks, stores clerks, and counter clerks. The data would seem to suggest that reading skills assessment should focus directly on the materials used on the job, and should evaluate the types of skills required for acceptable comprehension in the context of the specific job duties. It is not sufficient to evaluate "reading comprehension" in the abstract, or with the use of general models.

THE NATIONAL SURVEY OF READING ASSESSMENT PRACTICE

To explore the issue of how reading assessment practice matched the Smith model, I conducted a national survey, the report of which is attached as Appendix 3. A total of 63 jurisdictions returned the survey questionnaire. Of these, 44 (33 states, six cities, one county and four special jurisdictions) explicitly assess reading skills as a part of their employee selection program.

All but one jurisdiction use actual job materials as the source for at least some of their reading tests. The one remaining jurisdiction (along with 39 others) uses general purpose materials (e.g., textbooks, journals, etc.) which are related to the work of the target class. In addition five states, one city and two speciel jurisdictions use source materials not explicitly related to the target job. The most common "outside source" were newspapers and scientific/technical magazines. Only one state uses a variety of such "outside sources" for their reading tests.

All 44 jurisdictions attempt to assess literal comprehension, including the ability to follow directions, and the ability to locate specific facts or details in a selection, and three-quarters attempt to assess the ability to understand specific vocabulary.

All 44 jurisdictions attempt to assess some aspects of interpretive/ evaluative comprehension. All jurisdictions attempt to evaluate the candidates' ability to identify the main idea of a selection, and the ability to draw appropriate inferences based on information provided in a selection. Most jurisdictions (29 or more) also attempt to assess:

- o Ability to identify supporting/subsidiary ideas
- o Ability to identify supporting facts/relationships
- o Ability to understand the sequence of events

- o Ability to distinguish between cause and effect
- o Ability to identify similarities, dissimilarities, contradictions
- o Ability to properly classify or group facts, relationships, ideas
- o Ability to follow a line of deductive reasoning
- o Ability to follow a line of inductive reasoning
- o Ability to understand the point of view or purpose

"Syntactic Comprehension" was defined as the ability to correctly understand and use proper English syntax, usage and conventions in comprehending text. Somewhat over half of the 38 jurisdictions attempt to assess syntactic comprehension, with all of them trying to assess the use of both conditionals and transitionals. About a third try to assess the ability to identify temporal states and the ability to correctly identify antecedents and other referents.

The survey seems to suggest that national practice in large measure conforms at least superficially to the Smith model. Virtually all of the those explicitly assessing reading skills are drawing their stimulus materials from the materials used on the job, and many of the significant concepts sketched out by the questionnaire do receive heavy affirmation by the respondent jurisdictions.

However, the patterns of responses in the individual questionnaires, the supplemental materials enclosed with some questionnaires, the lack of reportable research and the narrative comments appended to the survey forms suggest that the coverage of the specific skills tested is largely a matter of chance. There is a clear suggestion in the responses that the respondent jurisdictions recognize in the questionnaire's categories concepts that they find in the test items they are using, but that the items were not consciously constructed to evaluate those specific skills, and certainly were not as part of any "map" of a defined domain. Certainly the survey shows only scattered attempts to evaluate syntactic comprehension, and attempts to evaluate interpretive/evaluative comprehension seemed to have major gaps.

The one item in the survey that I find interesting in the context of Smith's theory and the other information I have considered above is the use of cloze tests. Fifteen jurisdictions report the use of a multiple-choice versions of cloze tests. Given proper selection of stimulus materials, cloze tests by their very nature should come closer to mapping the set of conventions embodied in job-related text.

While I have not found any data that relates cloze results to a carefully mapped skill domain of the type suggested by Smith's work, Thomas Sticht, in his book *Reading for Working* says, "Research has indicated that, although there is no single definitive method for measuring reading comprehension, the 'mechanical' cloze procedure has consistently yielded very high correlations with multiple-choice tests and other more subjectively constructed measures of comprehension and difficulty.

"Therefore the weight of the evidence indicates that the cloze test provides a *valid* measure of reading comprehension. The fact that it is also strictly objective, and that *n* independent alternative forms can be created simply by deleting every *n*th word counting from the first, second, third,..., or *n*th word from the beginning of the passage, further encouraged the use of the cloze procedure."

Data from England on multiple-choice forms of cloze tests show KR_{20} reliabilities in the mid-0.90s, while studies on the *Degrees of Reading Power*, a multiple-choice form of cloze used widely in educational assessment, show similar KR_{20} reliabilities, as well as parallel forms reliability ranging from 0.83 to 0.91. These studies also found both convergent and discriminant evidence for the construct validity of the test.

An interesting finding of the DRP studies is that the multiple-choice cloze test is "Culture Fair": Students who received the same DRP scores acquired the same amount of information from ordinary prose regardless of their sex, ethnic background, or socioeconomic level.

CONCLUSION

I clearly find Smith's work compelling, and I find support for it in the work done on the two American corpuses, as well as in the data I find in my own tests. While I find the prospect of creating job specific reading skills tests daunting, I am beginning to think that some variation on the cloze procedure may be the way to address the problem.

APPENDIX 1 - TEST DATA

Item #	IRI	P		A	B	C	D	Omits
01	.210	.58	Hi	91	213	1013*	32	0
			Lo	362	307	557*	122	2
02	.175	.55	Hi	322	38	943*	46	0
			Lo	564	113	553*	113	7
03	.175	.73	Hi	23	1193*	19	114	0
			Lo	60	772*	138	379	1
04	.172	.69	Hi	152	1110*	25	62	0
			Lo	352	753*	110	129	6
05	.222	.66	Hi	120	56	32	1141*	0
			Lo	364	246	104	635*	1
06	.208	.77	Hi	82	15	3	1248*	1
			Lo	340	155	32	817*	6
07	.206	.32	Hi	678*	325	168	178	0
			Lo	185*	510	246	402	6
08	.249	.67	Hi	32	23	1188*	106	0
			Lo	231	173	617*	322	7
09	.161	.81	Hi	3	1254*	4	88	0
			Lo	19	920*	58	348	5
10	.159	.84	Hi	5	28	22	1294*	0
			Lo	49	167	148	980*	5
11	.190	.61	Hi	1020*	214	33	82	0
			Lo	613*	353	126	251	7
12	.170	.80	Hi	5	1260*	61	23	0
			Lo	87	901*	238	116	7
13	.194	.71	Hi	12	154	15	1168*	0
			Lo	149	359	87	749*	6
14	.184	.75	Hi	57	1200*	74	18	0
			Lo	195	822*	192	135	6
15	.265	.67	Hi	1206*	34	73	36	0
			Lo	591*	257	329	161	12

Mean - 10.144 SD - 2.940 KR$_{20}$ - .693 SE$_m$ - 1.629

Frequency Distribution N = 2699

15 - 103	11 - 363	07 - 157	03 - 33
14 - 222	10 - 296	06 - 143	02 - 15
13 - 320	09 - 281	05 - 86	01 - 2
12 - 359	08 - 249	04 - 67	00 - 3

APPENDIX 2 - SAMPLE TEST ITEM

DIRECTIONS: Each of the following 15 questions is related to the reading selection preceding the question. Base your answer to the question SOLELY on what is said in the selection - NOT on what you man happen to know about the subject being discussed.

"There is no simple, definable relationship between the amount of communication that takes place and the effectiveness of an organization. The advocacy of communication as essential to any organization is not enough. To be effective, an organization must concentrate on two specific areas: the kind of information required for the solution of its problems, and the nature of the communication process between individuals and groups at all levels of the organization. An organization must function as a restricted, focused communications network; unrestricted communication produces noise and inefficiency."

Which one of the following statements is best supported by the above selection?

A. The more communication there is between Individuals and groups within an organization, the more effective the organization is.

B. The importance of communications to an organization is dependent on the type of problems it must solve.

C. An organization's effectiveness is closely related to the informational content of its communications.

D. Lack of communication within the organization is the main cause of inefficiency.

APPENDIX 3 - READING SKILLS ASSESSMENT:

A Survey of Practice

Michael J. Dollard
New York State Department of Civil Service
Building 1, Harriman Office Campus
Albany, NY 12239

May 1989

INTRODUCTION

In January 1989 The New York State department of Civil Service a questionnaire on assessing reading skills to a variety of State, County, and City personnel offices, as well as to a number of special jurisdictions. A total of 63 jurisdictions returned the questionnaire. See Appendices 1 and 2 for the list of participating jurisdictions, and a copy of the questionnaire filled out with the aggregate responses of the 44 jurisdictions (33 states, 6 cities, 1 county and 4 special jurisdictions)that explicitly assess reading skills as part of their employee selection program.

In the following summary, specific data will be avoided as much as possible for the sake of clarity and intelligibility. The reader is urged to look at the display of aggregate data in Appendix 3A. Aggregate data only is presented in this report, consistent with the agreement made by the principal investigator with the participating jurisdictions. However, it can be said that the displays for the individual jurisdictions are generally quite similar regardless of the type of jurisdiction (i.e., State, county, city, special jurisdiction).

Of the jurisdictions that do not explicitly test for reading skills, several do not do any type of written testing, but rely on various types of evaluations of training and experience for employee selection. Others evaluate reading skills only indirectly. At least one jurisdiction in the latter category uses readability analysis to scale their written tests to the reading level required on the target job(s).

TARGET OCCUPATIONAL GROUPS

The survey showed the substantial use of explicit reading skills assessment in all occupational areas covered. Unfortunately, public safety positions were not listed in the questionnaire as a discrete group, so there is no separate data for them. Just less than half of the jurisdictions used explicit reading skills assessment for blue collar workers, and for managers/administrative workers. About two thirds of them use reading skills assessment for clerical workers and for professionals and technical specialists. Virtually all of the jurisdictions use reading skills assessment for paraprofessional and technical workers.

SOURCES OF READING TESTS

All agencies save one that do explicit reading skills assessment develop at least some of their testing materials in house. In addition, four states and one special jurisdiction use consultant-constructed test materials, while four states, two cities and one special jurisdiction use commercial "off-the-shelf" reading tests.

TYPES OF READING TESTS

The survey looked at three basic forms of test:

- CLOZE tests
- conventional written multiple-choice tests
- conventional open-ended written tests.

CLOZE tests are widely used in educational testing. In their classical form, a reading sample is selected, and then either every seventh or every ninth word is replaced with a blank. The candidate's ability to supply from the context the missing words is interpreted as an index of their reading ability. A variation on this classical form is a multiple choice format which provides four or five words or sets of words from which the candidate can select to complete the reading selection.

In the survey, two states and one city report using CLOZE tests in their classical form, while eleven states (including the two just mentioned), three cities and one special jurisdiction use a multiple-choice variation.

All 44 jurisdictions report using some form of conventional written multiple-choice tests, while virtually none of them report using any type of open-ended format.

SOURCES FROM WHICH TEST MATERIAL IS DRAWN

All but one jurisdiction use actual job materials as the source for at least some of their reading tests. The one remaining jurisdiction (along with 39 others) uses general purpose materials (e.g., text books, journals, etc.) which are related to the work of the target class. In addition five states, one city and two special jurisdictions use source materials not explicitly related to the target job. The most common "outside source" were newspapers and scientific/technical magazines. Only one state uses a variety of such "outside sources" for their reading tests.

LITERAL COMPREHENSION

All 44 jurisdictions attempt to assess literal comprehension, including the ability to follow directions, and the ability to locate specific facts or details in a selection, and three-quarters attempt to assess the ability to understand specific vocabulary.

INTERPRE'.IVE/EVALUATIVE COMPREHENSION

All 44 jurisdictions attempt to assess some aspects of interpretive/evaluative comprehension. All jurisdictions attempt to evaluate the candidates' ability to identify the main idea of a selection, and the ability to draw appropriate inferences based on information provided in a selection. Most jurisdictions (29 or more) also attempt to assess:

- Ability to identify supporting/subsidiary ideas
- Ability to identify supporting facts/relationships
- Ability to understand the sequence of events
- Ability to distinguish between cause and effect
- Ability to identify similarities, dissimilarities, contradictions
- Ability to properly classify or group facts, relationships, ideas
- Ability to follow a line of deductive reasoning
- Ability to follow a line of inductive reasoning
- Ability to understand the point of view or purpose

SYNTACTIC COMPREHENSION

"Syntactic Comprehension" was defined as the ability to correctly understand and use proper English syntax, usage and conventions in comprehending text. Somewhat over half of the 38 jurisdictions attempt to assess syntactic comprehension, with all of them trying to assess the use of both conditionals and transitionals. About a third try to assess the ability to identify temporal states and the ability to correctly identify antecedents and other referents.

TESTING TECHNIQUES

In the assessment of literal comprehension, virtually all jurisdictions use multiple-choice format requiring the candidate to identify appropriate facts in a selection in order to correctly answer a question.

In the assessment of interpretive/evaluative comprehension, half or more of the 44 jurisdictions use one or more of these multiple-choice formats:

- Recognize significant details contained in a selection
- Recognize the best paraphrase of a selection
- Recognize the best summarization of the material in a selection
- Apply explicit principles contained in a selection to a new factual situation
- Recognize the best statement of the author's point of view

RESEA?.CH

While 44 of 63 jurisdictions are explicitly assessing reading skills in one way or another, not a single jurisdiction reported that they have research reports they are willing to share. This does not necessarily mean that there are no studies, but simply that the studies are not for publication and distribution. It is expected that most of the studies -- as in the investigator's agency -- are "informal" studies done as a part of normal test development and administration, and are not in the format of a formal study. Indeed, many of these studies include the actual test material, which would be compromised by release.

*** APPENDIX 3A - QUESTIONNAIRE ON READING SKILLS ASSESSMENT ***

	42 states
Michael J. Dollard	11 cities
Principal Investigator	4 counties
(518) 457-2483	6 special jurisdictions
	63 total respondants

A. Local Testing Practices

1. Do you test specifically for reading skills? Yes(44) No(19)
[33 states, 6 cities, 1 county, 4 special jurisdictions]

2. For what job families?

a) Blue Collar Workers Yes(20) No( )

b) Clericals Yes(31) No( )

c) Paraprofessionals/Technicians (people with an AA level education) Yes(39) No( )

d) Professionals/Technical Specialists (people with BA and above) Yes(34) No( )

e) Managerial/Administrative Yes(23) No( )

3. Who is the test preparer?

a) Custom-prepared in-house Yes(43) No( )

b) Custom-prepared by an outside consultant Yes(5) No( )

c) Standard commercial test from a commercial vendor Yes(7) No( )

d) Other (Please describe)__________________________________

____

4. What kind(s) of Test(s) are used?

a) Cloze

1) Classical method - candidate must provide words to complete a selection (typically every 7th to 9th word is missing) Yes(3) No( )

2) Multiple-choice variation - candidate must replace from a list of provided words the words to complete a selection) Yes(15) No( )

b) Multiple-Choice

1) Find data in the selection to answer a factual question Yes(45) No( )

2) Recognize the best summary of a selection Yes(39) No( )

3) Recognize the best paraphrase of a selection Yes(32) No( )

4) Recognize the theme, perspective or central idea of a selection Yes(42) No( )

5) Recognize the best statement of the author's point-of-view or purpose Yes(31) No( )

6) Recognize the best supported conclusion leading from a selection Yes(38) No( )

7) Recognize the best supported classification or grouping of data in a selection Yes(19) No( )

8) Other item types (Please describe)____________________

____

____

____

c) Open-ended format (essay, oral, or short answer)

1) State the best summary of a selection Yes(0) No( )

2) State the best paraphrase of a selection Yes(0) No( )

3) State the author's point-of-view or purpose Yes(1) No( )

4) State the theme, perspective or central idea of a selection Yes(0) No( )

5) State the best supported conclusion leading from a selection Yes(2) No( )

6) State the best supported classification or grouping of data in a selection Yes(1) No( )

7) Other item types (Please describe)____________________

____

____

d) Other test types (Please describe)____________________

____

____

____

____

____

5. Source of stimulus materials

a) Job-related materials

1) Work materials (manuals, handbooks, rulebooks, etc.) taken from the work-site of the target class Yes(43) No( )

2) Work materials (manuals, handbooks, rulebooks, etc.) taken from work-sites of employees at a generally similar pay grade in either the public or private sector Yes(25) No( )

3) General purpose materials related to the work of the target class or occupational field (e.g., text books, journals, commercial manuals or handbooks for equipment, software or procedures used on the job Yes(40) No( )

4) Other (Please describe)____________________

____

____

____

Please continue to the next page.

b) Non-job-related materials

1) Newspapers Yes(9) No( )

2) Magazines

a> Popular press (e.g., *Readers' Digest*, *Redbook*, *People*, etc.) Yes(3) No( )

b> Literary press (e.g., *New Yorker*, *Harper's*, etc.) Yes(2) No( )

c> Current Affairs press (e.g., *Time*, *Newsweek*, etc.) Yes(3) No( )

d> Scientific/Technical press (e.g., *Scientific American*, *Omni*, etc.) Yes(10) No( )

3) Popular Fiction Yes(1) No( )

4) Popular Non-Fiction Yes(2) No( )

5) Other (Please describe)____________________________________

____

B. What reading tasks do you try to assess?

1. Ability to correctly understand and use proper English syntax, usage and conventions in comprehending text

a) Ability to correctly identify conditionals (e.g., if, when, should, etc.), and understand the relationship of the conditionals to the completion of the action Yes(27) No( )

b) Ability to correctly identify transitionals (e.g., but, yet, however, therefore, etc.), and understand the relationship of the transitionals to the completion of the action Yes(25) No( )

c) Ability to correctly distinguish between the actor and the object of the action (includes an understanding of 'active-' and 'passive-voice') Yes(15) No( )

d) Ability to correctly identify temporal states (past, present, future) Yes(19) No( )

e) Ability to correctly identify antecedents and other referents in written material Yes(22) No( )

f) Ability to recognize the use of figures of speech (irony, hyperbole, simile, metaphor Yes(10) No( )

2. Literal Comprehension

   a) Ability to locate specific facts or details in a selection Yes(44) No( )

   b) Ability to understand specific vocabulary Yes(36) No( )

   c) Ability to identify a person, place or thing based on a narrative description Yes(19) No( )

   d) Ability to follow written directions Yes(45) No( )

3. Interpretive/Evaluative Comprehension

   a) Ability to identify facts, ideas, and relationships

      1) Ability to identify the main idea Yes(45) No( )

      2) Ability to identify supporting and subsidiary ideas Yes(32) No( )

      3) Ability to identify supporting facts or relationships Yes(36) No( )

   b) Ability to understand the relationships among facts/ideas

      1) Ability to understand the sequence of events Yes(38) No( )

      2) Ability to distinguish between cause and effect Yes(29) No( )

      3) Ability to identify levels of hierarchy Yes(18) No( )

      4) Ability to identify similarities, dissimilarities and/or contradictions between facts, relationships or ideas Yes(32) No( )

   c) Ability to properly classify or group facts, relationships and/or ideas Yes(31) No( )

   d) Ability to draw appropriate conclusions based on information provided in written form

      1) Ability to apply the rules of argument (syllogistic reasoning) Yes(24) No( )

      2) Ability to follow (understand) a line of deductive reasoning Yes(35) No( )

      3) Ability to follow (understand) a line of inductive reasoning Yes(30) No( )

e) Ability to draw appropriate inferences based on information provided in written form Yes(44) No( )

f) Ability to formulate principles based on information provided in written form Yes(19) No( )

g) Ability to understand the point of view or purpose of a selection Yes(34) No( )

h) Other (Please describe)____________________

4. Other tasks to be assessed (Please describe)____________________

C. What methods of assessing reading skills do you use? (i.e, What is the task to be performed by the candidate?) [NOT APPROPRIATE FOR CLOZE-TYPE PROCEDURES)

1. Literal comprehension

a) Identify the appropriate facts in a selection to correctly answer a question Yes(40) No( )

b) Other Literal Comprehension methods (Please describe)____________________

2. Interpretive/evaluative comprehension

   a) Restate (open-ended) or recognize (multiple-choice) significant details contained in a selection Yes(34) No( )

   b) Paraphrase the selection (open-ended) or recognize the best paraphrase (multiple-choice) Yes(28) No( )

   c) Summarize the material presented in a selection (open-ended) or recognize the best summary (multiple-choice) Yes(32) No( )

   d) Apply explicit principles contained in a selection to a new factual situation Yes(25) No( )

   e) Apply implicit principles contained in a selection to a new factual situation Yes(18) No( )

   f) Complete an analogy based on information contained in a selection Yes(10) No( )

   g) State (open-ended) or recognize the best statement of (multiple-choice) the author's point-of-view or purpose Yes(28) No( )

   h) Other interpretive/evaluative methods (Please describe)______________________________________

   ____

   ____

   ____

   ____

3. Other assessment methods (Please describe)____________________

   ____

   ____

   ____

D. Do you have research studies related to the assessment of reading skills that you would be willing to share with other participants in this study? Yes(0) No( )

APPENDIX 3B - PARTICIPATING JURISDICTIONS

STATES (42)

Alabama
Alaska
Arizona
Arkansas
California
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Oklahoma
Oregon
Pennsylvania
South Carolina
South Dakota
Tennessee
Vermont
Virginia
Washington
West Virginia
Wyoming

CITIES (11)

Baltimore
Kansas City, MO
Los Angeles
New York
Phoenix
Pittsburgh
Portsmouth, VA
Roanoke, VA
Rochester, NY
Sacramento, CA
Toledo

COUNTIES (4)

Baltimore Co, MD
Henrico Co.,VA
King Co., WA
New Castle Co,
DE

SPECIAL JURISDICTIONS (6)

AMTRACK
Board. of Examiners, NYC Schools
Cooperative Organization for the Development of Employee Selection Procedures (California School Districts)
Port Authority of New York and New Jersey
U.S. Postal Service
U.S. Merit Systems Protection Board

BIBLIOGRAPHY

------ *Degrees of Reading Power: Description of a Reading Test and Its Compents* New York State Education Department (Division of Educational Testing), Albany: undated

Carroll, J. B., P. Davies and B. Richman. *The American Heritage Word Frequency Book*, Houghton Mifflin, Boston: 1971

Francis, W. N. and H. Kucera. *Frequency Analysis of English Usage*, Houghton Mifflin, Boston: 1982

Geva, Esther. *Cloze Reading Tests* (Published by Young, Hodder & Stoughton Educational, England) - Review in the *Ninth Mental Measurements Yearbook*, Buros Indstitute

Smith, Frank. *Comprehension and Learning: A Conceptual Framework for Teachers*, Holt, Rinehart and Winston, New York: 1975

Smith, Frank. *Understanding Reading: A Psycholinguistic Analysis of Reading and Learning to Read, 2nd ed.*, Holt, Rinehart and Winston, New York: 1978

Smith, Frank. *Psycholinguistics and Reading*, Holt, Rinehart and Winston, New York: 1973

Sticht, Thomas G. ed. *Reading for Working: A Functional Literacy Anthology* Human RTesources Research Organization, Alexandria, VA: 1975

THE MISSOURI SCHOOL SUPERINTENDENT ASSESSMENT CENTER CERTIFICATION PROGRAM

Richard C. Joines, President
Management & Personnel Systems, Inc.
602 Wiget Lane * Walnut Creek, CA * 415-932-2009

presented at the International Personnel Management Association Assessment Council (I.P.M.A.A.C.) conference, Orlando, Florida, 1989

OVERVIEW

Traditional assessment center 'echnology is oriented towards the identification of those candidates who are most qualified for promotion cr upward movement through management. Civil service systems routinely utilize the assessment center method to rank-order candidates and place them on civil service lists. As part of this process, candidates may also be failed either through a consensus decision by the assessor team or, as is more frequently the case in the public sector, by totaling the points obtained by the candidate to determine if it meets a prescribed total (or average) on the assessment dimensions. This paper describes the assessment program developed by the author for the certification of school superintendents in the state of Missouri and how the model differs from more commonly used assessment center procedures.

The State of Missouri is the first state in the country to pass a law requiring that school superintendents demonstrate their management skills in order to be certified. In addition, candidates must pass a multiple-choice knowledge test. In the first application of this model in April of this year, six candidates were tested in both processes. Two candidates failed the assessment center and one failed the knowledge test. These results have already sent a strong signal to potential certification candidates that certification is no longer a simple or routine matter. This paper will describe the assessment center model and the rationale for utilizing a non-traditional approach.

CONTENT VALIDITY CONSIDERATIONS/JOB ANALYSIS

It was clear from the outset that the State of Missouri desired an assessment program that could be implemented and defended without doing seveıal years of research. Moreover, it was not clear whether a criterion-related validation strategy was even feasible in their setting. In order to be so, it would have to be possible to obtain reliable and valid measures of performance of school superintendents. Since superintendents report to different Boards of Education across the state, consisting of lay people, it seemed unlikely that a method of evaluation could be developed that would reliably evaluate the performance of superintendents from one setting to the next. Thus, a content validity strategy was called for, and due to the high likelihood of litigation at some point, it was clear that this strategy needed to closely follow and emphasize the major content areas of the superintendent position.

A job analysis was conducted in which a committee of current school superintendents identified 66 tasks performed on the job. This committee also sorted tasks into duty areas, as follows: (1) curriculum & instruction; (2) personnel; (3) superintendent/board of education relations; (4) budget/finance; (5) community relations; (6) executive leadership; and (7) facilities management. A task questionnaire was developed a.ıd sent to all current school superintendents in the state (n=451). The tasks were rated on four variables: (1) time spent; (2) importance; (3) difficulty; and (4) initial performance.

Responses were received from 232 incumbent superintendents. A list of 45 "critical tasks" was identified by (a) excluding any task not required upon entry to the job; (b) multiplying the mean importance and difficulty ratings and adding the mean time spent; (c) identifying those tasks with a resultant value of 13.0 or higher and (d) including any task rated "4" or "5" by 50% or more of the raters. Interestingly, none of the tasks previously identified for the facilities management duty area met the criticality criteria; thus, this duty area was eliminated from further consideration.

The 45 critical tasks were used as the basis of a managerial skills questionnaire which was also mailed to all current school superintendents and 178 responses were received. This questionnaire supplied a list of 16 assessment dimensions for review by the raters. For each task, raters were asked to link the dimension to the task if the skill represented by the dimension was important to successful task performance. After completing the linkage process, raters provided a 1 - 5 rating of the importance of the dimension to overall success as a superintendent. Based on the questionnaire results, eight skills were identified for the assessment center, as follows: (1) oral communications; (2) leadership; (3) interpersonal relations; (4) problem analysis; (5) judgment & decision making; (6) planning & organizing; (7) management control; and (8) written communications.

LINKAGE OF JOB ANALYSIS TO EXAM PLAN

Content validity is based on representatively sampling the important requirements of the job. All too often, however, an extensive job analysis is completed and there is no apparent linkage to the actual exercises or exam content. The approach used in this study was to base the exam content on the specific duty areas and critical tasks contained in each. This insured a strong linkage between the content of the exercises and the job analysis, thereby insuring the content validity of the assessment center as a whole. For example, the following tasks were identified as the "critical" tasks in the superintendent/board of education relations duty area:

1. Prepares an agenda for each meeting of the board along with related and appropriate informative materials.

2. Communicates with board members to keep them informed and abreast of ongoing issues.

3. Meets with the board to discuss, review and resolve a wide range of issues.

4. Interacts with board members to mutually define respective roles and responsibilities of the superintendent and board of education.

These tasks resulted in the development of a simulated board of education meeting in which candidates must: (a) review letters from board of education members regarding several current issues; (b) prepare an agenda for the board meeting; (c) summarize and react to board member concerns on a number of the agenda items; (d) interact with board members who have varying personalities, interests and immediate concerns; and (e) attempt to develop cohesiveness and common direction among board members where these are currently lacking.

As may be seen, the "exercise" tends to closely parallel the tasks identified as critical in the job analysis for this duty area. The "simulation" is accomplished using five role players and explicit instructions on "what to say", "when to say it", and "how to react to all possible avenues" open to the candidate. Although it may not be easy, administratively or technically, to accomplish such a simulation, fidelity with job analysis results required such an approach.

This plan was followed for all duty areas identified in the job analysis. In order to representatively and comprehensively sample the critical tasks in each duty area, it was necessary in several duty areas to develop more than one simulation, resulting in a total of eight assessment exercises administered over the course of a two-day assessment center. In addition, the General Management In-Basket (GMIB) School Administrator version was used to assess generic management and administrative skills across duty areas. The assessment plan is summarized below.

DUTY AREA	EXERCISE(S)
Curriculum & Instruction	Curriculum and Instruction Analysis and Oral Report

DUTY AREA (continued)	EXERCISE(S) (continued)
Personnel	In-basket emphasizing personnel issues + interview
Superintendent/Board Relations	Board of education meeting (role play)
Budget/Finance	Review of budget/finance information & oral report
Community Relations	School/community relations analysis and oral report Meeting with community leaders (role play)
Executive leadership	Leaderless group discussion Management Interview
Generic management skills	General Management In-Basket (GMIB)

RATE EXERCISES, DIMENSIONS, OR BOTH?

Assessment center candidates are typically scored in one of two ways: (1) ratings are assigned on dimensions and a required total is established for passing (e.g., mean dimension score = 3.0 or higher); or (2) assessors review all available information and ratings and clinically combine the information, using their own individual judgment, to assign an overall assessment center rating, which may include a rating of pass/fail, promote or don't promote, etc. Feedback to candidates is typically on a "dimension" basis, with the candidate being given a final dimension score (e.g., decision making), then behavioral information across exercises to explain why the assigned rating was made.

There is substantial evidence in the research literature on assessment centers to indicate that construct validity of dimension ratings is lacking (Sackett and Dreher, 1982; Klimoski and Brickner, 1987). Sometimes referred to as the "exercise effect", correlation of ratings within exercises are often stronger than correlations of the same dimensions across exercises. In the author's experience, this likelihood is stronger to the extent that the assessment model is thoroughly based on content validity. This should not be surprising in that the tasks performed by incumbents of any management position may differ significantly, thereby leading to situations in which incumbents are skilled in one area but not another. Thus, should their decision making be excellent in one instance, yet poor in another, with the likely explanation being difference in task content, should we conclude that the individual is "average" in decision making -- or excellent in one setting and poor in another?

The certification model herein described has included both sets of information for review by assessors in reaching final "pass/fail" decisions. A sample set of scores is given in Table 1. A strong "exercise effect" may be seen in the data by comparing score on dimension ratings in the budget/finance exercise with scores on the same dimensions in other exercises. These results indicate that the individual is deficient in budget/finance skills but not necessarily that he/she is a poor decision maker in other areas.

As may also be seen this candidate obtained an average dimension and exercise score of greater than 3.0 (1-5 scale) but was nevertheless failed in the assessment center as a whole. This resulted from deficient ratings in the three exercises which have been asterisked (Personnel I.B. + interview asterisked because of deficiencies in dimensions unique to the exercise, e.g., written and oral communications not unique). For this candidate to be certified, he/she will have to either take the entire assessment center again and pass or, as is more likely, successfully retake those exercises which resulted in the failure.

TABLE 1

SAMPLE ASSESSMENT CENTER CANDIDATE RESULTS

	LGD	MGT. INT.	SCHL COMM.	BOARD ED.	COMM. LEAD.	PERS I.B.	C & I	BUDG. FIN.	GMIB	SKILL RATING
ORAL COMMUN.	3	3	3	4	4	4	3	4		3.50
LEADERSHIP	4	4	3	4	4					3.80
INTERPERS. RELATIONS	3		4	5	5	3				4.00
PROBLEM ANALYSIS	3			4		3	3	1		2.80
JUDGMENT/ DEC. MAK.	3	4		4		2	2	1		2.67
PLANNING & ORGANIZING			3	3	4	2	3	1		2.67
MANAGEMENT CONTROL						3		1		2.00
WRITTEN COMMUN.						4		4		4.00
EXERCISE RATING	3.2	3.67	3.25	4.0	4.25	*** 3.0	**** 2.75	*** 2.0	*** 2.0	FINAL = 2

FAIL

OVERALL MEAN SKILL RATING = 3.180

OVERALL MEAN EXERCISE RATING (WITHOUT GMIB) = 3.265

OVERALL MEAN EXERCISE RATING (WITH GMIB) = 3.124

CANDIDATE FEEDBACK

The model of looking at exercise scores instead of solely at dimension scores has some important benefits related to feedback of results to candidates. Instead of making generalizations about underlying constructs, such as an individual's problem analysis or decision making skills, the feedback can emphasize the specific areas in which the candidate's skills are deficient related to a specific duty area of the superintendency. Beside being less threatening, it is also more consistent with assessment center research findings.

CERTIFICATION AS A POLITICAL ISSUE

The decision to engage in a program of certification is likely to engender strong feelings on the part of all who are involved. Educators apparently have mixed feelings about the subject. On the one hand, they tend to agree on the need for competency testing, but on the other, there is a feeling that local control is being usurped. Perhaps this is nothing unique to the field of education but a general characterization of the issues that are involved in a state government attempting to establish minimum standards across local government entities.

From the perspective of an outside consultant, and as general guidance to others who might undertake such a program, beware! The underlying currents of discord and mixed feelings may be quite strong. Systems which have been tested and proven to work in other settings may suddenly be questioned as the reality of the enormity of the career decisions that are about to be made are actually confronted.

After those decisions have been made, however, the tension is likely to subside provided that: (1) the assessment center has indeed been thorough; and (2) incumbents of the position, in this case current superintendents, have had a large degree of involvement and have bought into the process. Their support will carry 99% of the day in dealing with new and evolving political issues.

REFERENCES

Joines, R. (1986). Put superintendent candidates through an assessment center. The American School Board Journal, 173, 31 -32.

Klimoski, R. and Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. Personnel Psychology, 40, 243 - 260.

Sackett, P.R. and Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67, 401 - 410.

DEVELOPING THE WRITTEN SIMULATION: A CASE STUDY

by Judith Trabert, Ph.D
City of Rochester, New York

In 1986, the Civil Service staff of the City of Rochester, New York was faced with a problem. Two years before, an assessment center for a promotional police sergeant selection process had been developed. It had included an in-basket exercise, two roleplays (one with a problem employee and one a community relations situation) and a leaderless group discussion. The assessment center had been administered to the 50 candidates (out of about 100) who had scored the highest on a written multiple choice exam developed by New York State Civil Service. (Prior to this time the State exam had been used by itself to screen and rank candidates.)

The successful aspects of the 1984 assessment center had been high candidate acceptability and little adverse impact in contrast with the written multiple choice exam. The downside of the assessment center had been cost -- in time and dollars -- and what Civil Service felt had been a lack of objectivity among the internal Police Department raters. Rater training and exam administration had taken three weeks and had been a severe drain on the Department's command personnel. In addition, although care had been taken not to assign to a candidate any assessor who had supervised him or her, most police department personnel knew each other either personally or by reputation, and Civil Service thought that expectations for performance had been created in the assessors that training could not or had not erased.

So by 1986, a dilemma existed. The police department felt it couldn't afford to pay for assessors from outside the department, and Civil Service felt it didn't have the staff resources to develop and administer another full assessment center. So the City, advised by consultant Dr. Nancy Abrams, looked around for a testing methodology that would have the advantages of the assessment center without the problems. The assessment center exercises had been used to test for complex problem analysis and decision-making skills as well as oral communication and interpersonal relations. Inspired by the accounts of the work being done by New York State Civil Service with written simulations, Rochester devised a new test plan which included written simulation exercises, a short oral exam, and a new locally developed written test. Under the new plan, the written test would assess such knowledges, skills, and abilities (KSAs) as knowledge of criminal law and police procedures, English grammar, punctuation and spelling, and the ability to read, research and interpret written materials such as General Orders and the penal and vehicle codes. The short oral exam -- 1/2 hour per candidate -- would concentrate on oral communication skills -- listening, interviewing to obtain information, expressing oneself. And the bulk of the KSAs, such as:

Ability to evaluate the effect of a solution on a problem;
Judgment, the ability to draw a logical conclusion based on facts;
Ability to change direction or response to situations as they progress;
Analytical ability;
Ability to interact with groups, media representatives;
Ability to recognize a problem situation;
Ability to recognize and assess officers who are expressing personal or work-related problems;

were loaded onto the written simulation problems, either in the content of the problem or its structure. This was going to be a short, cheap assessment center, The Solution. It was, as you might expect, not quite that easy.

When Civil Service began the development process, neither analysts or subject matter experts (SMEs) had been trained in how to develop a written simulation. It was make it up as you go. Problems in five content areas were to be developed -- a problem employee, community relations, protecting a crime scene, and research and tactical problems. Each problem was assigned to a team which included two police SMEs -- Sergeants, Lieutenants or Captains -- and one analyst. The SMEs supplied ideas -- what would happen next in this situation, what alternatives would the sergeant have, what would be the consequences of each -- and the analyst would turn this raw material into a simulation. As the development process progressed, Civil Service learned that successful SMEs had good verbal and analytical skills, highly expert knowledge of the topic, and a certain amount of creativity and imagination. A successful analyst had good analytical and communication skills, and very importantly, was an expert logician. It was the analyst's responsibility to make sure that the problem was internally consistent. This meant that SMEs who went off on tangents or invented sequences that didn't fit the overall logic of the problem had to be pulled back to predict the long-term, real-world consequences of every decision.

And we found that not everyone could do it. Some analysts who were competent constructors of multiple choice items or single situation decision problems could not master the intricacies of the written simulation's logical loops. The best analyst at this turned out to be a former math teacher for whom a simulation path resembled a complicated geometry proof.

So the simulations were developed, over a five to six month period, and then it was time to get them printed. A longer-than-anticipated development timeline meant that the exam was now only a few weeks away. Commercial printers who did latent image printing were very scarce, and their schedules called for much more lead time than Civil Service had. So instead they resorted to a wax rub-off printing technology similar to that used for lottery tickets. For some reason - perhaps insufficient drying time - the wax cracked and peeled, and exam pages stuck together. Exam administration day was a logistical circus. But, in spite of the production and administration problems, the examination was judged a success.

The new examination plan had turned out to have all the advantages of the assessment center. It had less adverse impact than the written test, high acceptability to candidates, and produced a list which was in line with the Department's judgment of its more or less qualified potential sergeants. In addition, the written simulation had been administered in one day, and scored by Civil Service personnel rather than Police. The drawbacks of the written simulation had been extensive development time and the need for especially skilled analysts on the development team. All in all, the written simulation was judged a success, and the team geared up to develop an alternate exam for the next administration with enthusiasm and ideas for improvement.

USE OF A WRITTEN SIMULATION TEST AS AN EXAMINATION PART-ALTERNATE FORM EXPERIENCES

Nancy E. Abrams
Consultant, Personnel Management
and Measurement

Judy Trabert's paper has described the development and use of a written simulation test as part of the examination for Police Sergeant in Rochester, New York. In the second process of test development, we attempted to develop an alternate form to the original examination.

Development The test development process began with a review of the test content. We assembled those who worked on the previous exam to work on the new one. To the extent feasible, this included the same subject matter experts and Personnel Technicians. The new simulations were developed to cover the same content areas as the old ones, that is a Problem Employee, Protection of a Crime Scene, Community Relations, Tactical, and Research problems. The SME's who worked on the problems were given copies of the old problems and were asked to change the "story" but not to significantly change the decision-making process. In cases where the same SME's worked on the same problems as the previous examination, the process was accomplished reasonably efficiently, although there was significant SME burnout. In cases where new SME's were working on the problems, they sometimes had difficulty with their task. They needed to study the problem and understand its basic logic. They found this difficult. The development of the alternate form version of the test was accomplished more rapidly than the original test, but of course we also knew more.

Results The new test was given as part of the examination process. Because of some of the administrative problems Judy mentioned, we changed the test part order. In the earlier test we gave the written knowledge test, then the simulation and then an oral. The second time the simulation and the oral were reversed. A multiple hurdles approach was used with few failures on each part. The tests were designed, to the extent feasible, to test for different KSA's.

The overall results appear in Table 1. As you can see the test parts were only modestly correlated with each other, indicating that they appeared to be testing different KSA's. For some reason the intercorrelation of test parts increased with the new test. We can see that for those who took both the new and old test, the scores on the oral were quite consistent. This was much less true for the knowledge test and the simulation. Although this is a small sample, it seems to indicate a low level of alternate form reliability. But is it?

We also compared the parts of the test. Tables 2A and 2B show these results, first for the scores which had been converted to a seven point scale and also for raw scores. These results indicate essentially zero order correlations for the two sets of test problems. Means were somewhat lower and standard deviations slightly larger for the new test.

Discussions

These results seem to indicate that we failed in our attempt to develop an alternate form of our test. The data is limited and I am not sure that is what we found. Oral exam scores were reasonably consistent. The oral exam was designed primarily to test basic oral communication skills. These are not really likely to change over the course of the two years between the two tests. The written knowledge test score can certainly be changed through study. The low correlation for this test may indicate that those taking the new test, generally had not scored highly on the earlier test and had not been promoted, were motivated to study for the new test. The same phenomenon may have been true for the simulation.

I might argue that the learning factor may play a larger role for written simulations. Often we end the candidate's test by communicating success or failure in the result. This might lead candidates to seek out the correct answer and certainly may teach them that their choices were wrong. The written simulation test may actually be a teaching tool itself. If this is true, this may present us with some real problems in developing alternate form simulation tests. I hasten to caution you that these results are based on a very limited data set. We need to conduct further studies with careful attention to these findings to know whether our hypotheses are correct or not.

MED-TOX
ASSOCIATES, INC

Empirical Validation of Firefighter Vision Standards: Four Field Experiments in Montgomery County, Maryland

Vernon R. Padgett, Ph.D.
MED-TOX Associates, Inc.
Tustin, California 92680

I. Introduction

The National Fire Protection Agency requires firefighters to have uncorrected distance acuity of at least 20/40 in the better eye. OSHA standards, adopted from ANSI standards written in 1970, forbid the wearing of contact lenses under full-face respirators because "they could be blown out of the eyes." Neither standard was related to the job via job analysis and both were considered too strict by firefighters. Because of widespread concern about the legitimacy of these vision requirements, an empirical validation of the distance vision standard was planned.

II. Need for Vision Standards

The need for an uncorrected standard, that is, a standard for vision in the absence of corrective lenses, is determined by the need for the firefighter to do the job safely on those occasions when corrective lenses are missing.

When setting the standard, two demands must be balanced: 1) the need for safety and protection of firefighters and the public, and 2) the avoidance of unfair discrimination against applicants who can perform the job adequately, even though they lack perfect vision.

There is little empirical ground on which to justify any particular firefighter vision standard. Firefighters vision standards appear to have been based largely on opinion survey from very small samples of ophthalmologists or have been borrowed from the police or the military. Not only do firefighters vision standard vary widely for uncorrected vision, but vision standards in general show little agreement across regions: Holden's (1984) survey of 323 police departments showed that some required 20/20 uncorrected vision whereas [illegible] percent set no uncorrected standard at all.

III. Transportability

It appears that vision standards for firefighters have not been based on formal analyses of visual demands in firefighting. Transporting police vision standards to firefighters is inappropriate.

IV. Survey and Records Review

The purpose of this effort was: 1) Establish a need for a standard; 2) Provide critical far visual task information; 3) Assess the similarity between responses of Montgomery County contact lens-wearing firefighters and those described in the two 1985 survey reports; 4) Allow comparisons of vision obstruction between all possible pairs of firefighters who a) wear glasses; b) wear contacts; c) don't wear corrective lenses; and 5) Describe the population of Montgomery County Firefighters

V. Results of Records Review

Review of Insurance Program listing worker's compensation cases 1983-1986, established a number of cases in which firefighters worn glasses, they would certainly have not been in place following injury.

VI. Results of Survey

The survey results corroborated results of two earlier reports on soft contact lens use: Kartchner's (1985) experiment and the da Rosa and Weaver (1985) survey. Response rate: Of the 814 career and approx 800 volunteer firefighters, 525 returned surveys. Response rate for career firefighters was 80 percent. Response rate for volunteers was 21 percent. Need for a vision standard is formally established.

VII. Critical Vision Task Analysis

Critical firefighters duties were assessed using the Physical Abilities Analysis Manual developed by Carmean in 1983 and used by Hogan and Fleishman in their later work. Survey responses to an item on critical vision tasks supplemented responses collected during job analysis. Examples of critical vision tasks: "Identify a fire victim in a residence fire"; "Identify hazardous materials sign on truck." In all, 23 critical vision tasks were identified.

Relative criticality was assessed through ratings obtained from experienced first-line Montgomery County glasses-wearing firefighters on the following scale:

```
 Doesn't    []-----[]-----[]-----[]-----[]-----[]-----[]   Consequences are
  Matter    1      2      3      4      5      6      7     Multiple Deaths
  at all
```

Ratings allowed an empirical determination of the most important tasks.

VIII. Simulation Construction

Critical vision tasks were selected on five dimensions in order to provide a basis for the development of the simulations: 1) Criticality-- how important is the task?; 2) what is the probability of performing this task without glasses (or contacts); 3) Can performance on this task be objectively scored?; 4) Can this task be realistically simulated?; and 5) Is this task independent of learning or job experience? The five criteria were applied to the 23 critical far vision tasks. The four tasks selected for simulation were based on these selection factors. These were:

1. Identification of hazardous materials signs on a tanker.
2. Identifying dangerous materials across a room during overhaul after a building fire.
3. Spotting a victim on the fourth floor of a building.
4. Spotting a victim behind a window on the fourth floor of a building.

These simulations were fairly straightforward translations of those tasks judged most critical.

IX. Determination of Sample Size: How many Ss?

I determined the needed sample size to reject the null hypothesis with a pcwer of .90 and alpha - .05 after determining the effect size of the independent variable. I did this by examining Giannoni's anova on his 1981 report on CHP vision standards validation (a research program which provided a model for our methodology). The Eta-squared from Giannoni (1981) corresponded to a population correlation of .95. Based on Cohen and Cohen (1975), a sample of four subjects would provide the desired power of .90. I chose a sample of six to be safe.

X. Experimental Assessment of Visual Acuity Requirements: Procedure

Participants were six Montgomery County career firefighters self-identified as having 20/20 vision. All had between 5 and 13 years first-line firefighter experience. Age ranged from 21 to 31. Each firefighter was decorrected during a standard ophthalmological exam to four levels of visual acuity ranging from 20/40 to 20/200 and was provided with four pairs of glasses. Choice of levels of decorrected acuity was based on previous research, professional opinion, and Montgomery County regulations. Clinical decorrection of 20/20 sighted subjects represents a conservative approach to visual acuity standards validation.

Firefighters were tested individually in a repeated measures design on the four simulations in a single day. The head experimenter tested each firefighter's uncorrected vision, and tested him again with each of his four pair of glasses to verify decorrection. Glasses were clearly labeled with their decorrection. Previously designed procedural steps and continual monitoring during the experimentation controlled for squinting, peeking, and for wearing correct glasses. Two experimenters independently scored firefighter performance on the simulations; interrater agreement was 100 percent.

XI. Experimental Assessment of Visual Acuity Requirements: Results

The levels of performance were obtained at each visual acuity level in each simulation. For example, mean adequate performance in Simulation 1 occurred at

184 feet with 20/20; 155 feet with 20/40; 106 feet with 20/70; 80 feet with 20/100 and 47 feet with 20/200 vision. Anova on these data is given here:

Analysis of Variance Source Table for Results of Simulation 1:
Identification of Hazardous Materials Warning Signs

Source of Variance	df	SS	MS	F	p	Eta^2
Acuity (Columns)	4	73,889.5	18,472.4	56.13	.001	.89
Firefighters (Rows)	5	2,411.4	482.3	1.47	.25	.03
Residual	20	6,582.5	329.1	-	-	-
Total	29	82,883.3	-	-	-	-

Overall Results: 20/100 vision provided adequate performance on two simulations and 20/200 acuity allowed for passing performance on the other two.

XII. Setting Standards

The expert panel, composed of fire chiefs, fire training officers, and an applied psychologist, had determined acceptable cutoff scores for adequate performance for each simulation prior to the conduct of the field experiments. For Simulation 1, 80 feet was judged acceptable as a minimum distance from which firefighters should be able to identify hazardous materials signs and perform other tasks requiring similar identification. (This distance incorporates considerations of average street setback from most Montgomery County buildings, congested traffic conditions, and the distance from a potentially hazardous tanker judged safe by fire experts.) Minimal performance levels were similarly derived for Simulations 2, 3, and 4. Based on firefighters' performance on each simulation, an overall binocular far vision standard of 20/100 was recommended.

XIII. Contact Lens Issues: Overview of Survey Results

Two 1985 surveys cited above showed no reason to restrict firefighters from soft contact lens wear. Similar findings were reported in survey research of Los Angeles police conducted by Drs. Staley and Goldberg (1988); their findings noted that glasses produced more problems than contacts. Our survey of 40 firefighters corroborated these previous reports.

XIV. Ancillary Issues

A. Aging and acuity decrement
B. Protective Glasses
C. Vision obstruction reported by glasses-wearers versus contact lens wearers versus no corrective lens group
D. Generalizability of findings beyond these four simulations (Briggs)

XIII. References

168

THE DEVELOPMENT OF TRTs

Richard F. Thornton
Educational Testing Service
Princeton, NJ

Tailored Response Testing (TRT) is a new type of test that was developed at Educational Testing Service (ETS). It has demonstrated its applicability to the evaluation of human performance in a wide variety of occupations and work settings. This paper sketches out the conceptual and early developmental aspects, and discusses the application of the methodology to several types of jobs.

The TRT was designed to replicate more closely than do most other types of tests the functions and conditions encountered in occupations. The TRT retains a number of the realistic aspects of the job, assesses critical job functions, and is inexpensive to administer and score.

The TRT consists of two parts: (1) the stimulus providing a situation and (2) text that the examinee is to tailor or edit. A variety of stimuli can be used, e.g., videotapes, depicting actual or staged work activities, replicating job content and conditions commonly encountered on the job.

The examinee edits text which might describe: interpretations of events in the scene shown; evaluation of the effectiveness of actions taken; descriptions of corrective measures; and indications of additional information needed to resolve the problem. The examinee edits the text by crossing out words, phrases or whole sentences, so that the final text represents his or her judgment about the information presented.

An examinee's score is based upon his or her crossing out words that should be removed from the text, minus those words crossed out that should not be deleted. One advantage of this response mode is that it provides few cues as to the location of the words to be crossed out. Examinees must draw on their knowledge of and experience in the field to recognize and change inappropriate courses of actions and to know what needs to be done under the circumstances described.

Research has shown that the TRT method yields results very similar to those obtained from open-ended response simulations with no loss in predictive power against independent, work-related criterion measures.

This paper also discusses methods for developing TRTs that meet the requirements of the Guidelines for Employee Selection Procedures.

TECHNIQUES FOR CONSTRUCTING TRTs

Stanley J. Kalisch, Jr.
Educational Testing Service
Atlanta, GA

This paper addresses the technical concerns encountered in applying the TRT technique to an actual assessment problem. It details the general developmental steps involved, and discusses some of the problems that may be encountered as well as some of the precautions that must be taken.

Various stimuli have been used for TRTs. Video-taped presentations of scenarios have been the most commonly used stimulus. Video presentations provide many advantages over other forms of stimuli, e.g., information via the printed page, or static displays such as video slides. Motion often provides subtle information that the test developer wants presented, without conspicuously calling attention to particular details. Multiple stimuli can be used with any single TRT item.

The construction of the text for TRTs requires careful attention to the ways in which the text could be edited. The test developer must avoid multiple correct ways of editing the text. As with all item types, the developer must be cognizant of item constructions that give the examinees clues to the correct answers; fortunately, the TRT response mode offers advantages in this area of concern.

Because the TRT item type is new to the assessment field, rules guiding their construction are evolving. Lessons learned in the development of this item type will be discussed.

An actual practice item used in the TRTs developed for the Navy will be shown on the video cassette recorder and used to demonstrate the precise method used to administer, take, and score the TRT items.

APPLICATIONS AND RESULTS OF TRT

Herbert George Baker
Navy Personnel Research and Development Center
San Diego, CA

This paper includes a discussion of the context in which TRT has been applied to Navy jobs, shows video segments of actual job content for three jobs, demonstrates an actual test item from the Fire Controlman test, and presents the score distribution from a TRT administration to Electronics Technicians. Thus, it addresses the results of TRT application and actual testing in the real work environment of the Navy.

Recently, TRTs were developed by ETS for three Navy jobs (called ratings). These ratings were: (1) Electronics Technicians (ET); (2) Fire Controlman (FC); and Gas Turbine Systems Technician - Mechanical (GSM). The TRT created for the FC rating actually required the development of two tests because of the intra-job specialization into radar and data operators. (Participants viewed video segments of actual job tasks in each rating.)

These TRTs were developed within the Navy Job Performance Measurement (JPM) Program. This multi-year effort is aimed at linking selection tests with on-the-job performance within the military, and each Armed Service has a similar project under way. In the JPM Program, hands-on (or job sample) tests are being developed for each of eight Navy ratings. In addition, for each rating, one or more surrogate (substitute) tests are being developed and evaluated. TRT is one of these potential surrogates in at least three of the eight ratings.

Each of the TRTs developed for the Navy use video-presented material as the main stimuli, supplemented in several cases with some brief auxiliary printed materials. The test materials include a video cassette recorder (VCR), a videocassette, a set of printed responses (test items), a supplementary materials and notepad, and a manual of administration. The Navy TRTs were administered at several test sites throughout the United States, to personnel from the relevant ratings. Seminar participants had the opportunity to view an actual FC test item, and try their hand at TRT. In addition, descriptive statistics are presented for TRT results on a sample of 120 first-term ETs.

SUMMARY

CRITERION VALIDATION OF A PREEMPLOYMENT PSYCHOLOGICAL TEST FOR CORRECTIONAL OFFICERS.

INTRODUCTION

The hiring of personnel for the typical Correctional Officer position has always been a sensitive and difficult issue. Failures in the hiring process in this area are particularly visible. This has been the case in Iowa, as well as other jurisdictions.

Iowa law requires that a psychological test be used to assist in preemployment screening. The Minnesota Multiphasic Personality Inventory had been used for that purpose, but had not been validated specifically for use in Iowa. In addition, procedures for use were less standardized than desired, and interpretation for use in an employment setting was difficult. Concern over administration problems spurred by several incidents of apparent failures on the job, and resultant liabilities, caused this study to be undertaken.

This study was designed with the assistance of Dr. John T. Flynn of Flynn Associates of Hampton, Connecticut. Actual conduct of data analysis was conducted by Dr. Flynn and his associates, but the actual data gathering and administration of the project were carried out by staff from the Iowa Department of Personnel.

APPROACH

Because of the very nature of the use of a psychological test, it was evident that a criterion validation study would be required to insure that use of such a test would be appropriate. Procedures used in this study were designed to meet professional standards as expressed in the A.P.A. guidelines and the Federal Uniform Guidelines on Employee Selection.

A job analysis was conducted to identify behaviors judged by subject matter experts to be associated with the positions of Correctional Officer and Senior Correctional Officer. The behavior statements and background information gathered were used to develop a rating instrument for use with a sample of 362 Correctional Officers and 50 Senior Correctional Officers using a Likert type rating scale designed to measure the frequency and importance of these behaviors.

Information from an analysis of data from these ratings was used to develop the actual performance rating devices used as criterion references in this study. Performance rating devices were phrased as behavior statements. Ratings were, again on a Likert type scale, designed to measure the frequency that behavior was displayed by the Officer being rated. Performance ratings for both groups were reliable. (.98 for Correctional Officer and .96 for Senior Correctional Officer) and were normally distributed.

The California Psychological Inventory (CPI) and the Correctional Officer Interest Blank (COIB) tests were administered to a sample of 280 Correctional Officers and 28 Senior Correctional Officers. Regression analysis was applied to the score on all 20 scales of the CPI and the score on the COIB and the performance evaluations. This yielded a Multiple R of .3980 and an R Square of .15841.

Dr. Flynn then performed a Setwise Multiple Regression to identify a possible subset of COI scales and the COIB score to see if prediction could be increased. Results showed that the COIB score and scores on five scales of the CPI approximated the Multiple R of the total score. The Multiple R was .39059. Scales used included Psychological Mindedness, Dominance, Tolerance, Empathy and Responsibility. This was then cross validated. The resulting beta weights were inserted into the formula for predicted score as follows:

Predicted Score = 138.03 + (.26 X COIB) + (-.208 X PSYMIND) +
(-.164 X DOM) + (.23 X TOLERATE) + .026 X
EMPATHY) + (.069 X RESPON)

The correlation between the predicted score and actual job performance score was r = .3839 and it was st tistically significant.

Results for the Senior Correctional Officer were similar. There was some difference in scales on the CPI. The final prediction formula was:

146.05 + (.496 X COIB) + (-.061 X RESPON) + (-.166 X INDEN) +
(.376 X TOLER) + (-.343 X PSYMIND) + (.181 X DOM)

Actual Multiple R for Senior Officers was .62. A Multiple Discriminant Function Analysis performed for the Correctional Officer sample showed a 70.8% predictability of group membership and a 71.4% predictability rate for Senior Officers.

CONCLUSION

This study showed that certain scales of the CPI could be combined with the COIB score to serve as a valid measure of future performance for both job classes involved in this study.

The CPI and the COIB are both currently being used as one part of a comprehensive selection system for these classes. The prediction formulas provided by Dr. Flynn are being used as the basis for providing a final score for Correctional Officer and Senior Correctional Officers. Currently a low passing point of a obtaining a score in the top two-thirds of the predicted performance range is being used until more data can be gathered. These tests are being hand scored currently, with all selection information being entered onto a central data base. This data base calulates the final psychological test score, and includes it with all selection information on applicants. This information is available online to all Corrections Department institutions. A follow-up study of the relationship actual portions of the preemployment process have to job performance of applicants currently being hired is being planned for the future.

BIBLIOGRAPHY

Achen, C.H. Interpreting and using regression. London. A Sage University Paper. 1982.

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1985). Standards for educational and psychological testing. Washington.

Flynn, J.T., Flynn, B. E., & Bartolini, R. Understanding Psychological Test Scores. Hartford, CT. Barker Press, 1982.

Fowler, Floyd J., Jr. Survey Research Methods. Sage Publications. Beverly Hills. 1984.

Gough, H.G. Manual for the Correctional Officers' Interest Blank. Palo Alto, California: Consulting Psychologists Press, 1982.

Gough, H.G. Manual for the California Psychological Inventory. Palo Alto, California: Consulting Psychological Press, 1987.

Hathaway, S. R., & McKinley, J. C. Minnesota Multiphasic Personality Inventory, revised manual. New York: Psychological Corporation, 1967.

Lewis-Beck, M. S. Applied Regression: An Introduction. London. A Sage University Paper. 1982.

Sumary/cr

David Lundquist
Iowa Department of Personnel

Decision Making in Assessment Centers

Ira T. Kaplan, Arthur Kramer,
and William Metlay

Hofstra University

Previous research on group decision making has shown that overloading the group members with information decreases the accuracy of their judgements (Streufert & Driver, 1963; Streufert & Schroder, 1965; Streufert, Suedfeld, & Driver, 1965; Streufert, Clardy, Driver, Karlins, Schroder, & Suedfeld, 1965). Studies of consumer choice behavior suggest that the number of alternatives considered has a larger effect on decision accuracy, than does the amount of information provided about each alternative (Jacoby, Speller, & Berning, 1974; Jacoby, Speller, & Kohn, 1974).

The present experiment examined the effects of information load on group decision making in the context of personnel selection in a simulated assessment center. The alternatives were candidates for a managerial position and the information consisted of test scores for each candidate. Two dimensions of information load were manipulated: (1) the number of candidates considered and (2) the number of scores per candidate. Dependent variables were group decision accuracy and member attitudes.

Method

Subjects

The subjects were 24 student volunteers from undergraduate courses in industrial and organizational psychology.

Procedure

Six mixed-gender groups of four members each ranked candidates for a management position. The selection procedure was based on that used in a management assessment center, as described by Sackett and Wilson (1982). The design was a 2 x 2 randomized factorial, in which the independent variables were number of candidates (3 or 12) and number of performance dimensions (3 or 12). Each group was exposed to all four experimental conditions in counterbalanced order.

Measures

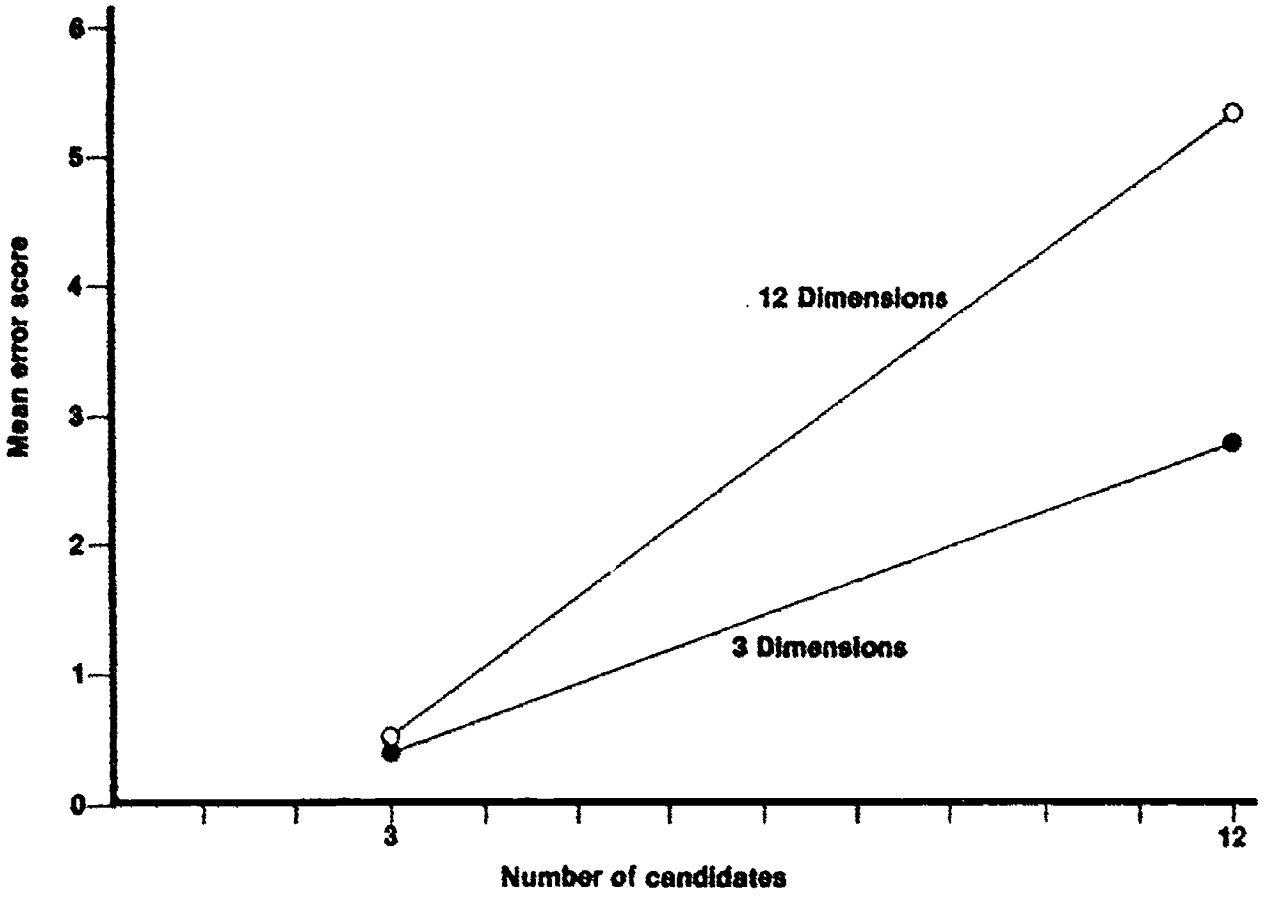
Decision accuracy was determined by comparing the group's ranking of the candidates with the optimal solution. In addition, the group members' attitudes toward their decision were measured after each problem by means of a questionnaire that assessed satisfaction, confidence, confusion, desire for more or fewer candidates, and for more or fewer test scores.

Results

Accuracy

As shown in Figure 1, groups made few errors when they ranked 3 candidates on either 3 or 12 dimensions. They made more errors when they evaluated 12 candidates on 3 dimensions, and still more with 12 candidates and 12 dimensions.

Figure 1. Effect of information load on decision accuracy.



174

Attitudes

Information load had no significant effect on confidence or confusion. Group members indicated that they were generally confident and not confused under all conditions.

Satisfaction decreased significantly (although the subjects were still more satisfied than not) when the number of candidates was increased from 3 to 12, regardless of the number of dimensions.

Varying the number of candidates had a significant effet on the desire to work with more or fewer candidates. After evaluating 3 candidates, the group members indicated a desire for more candidates; after evaluating 12, they expressed a preference for fewer candidates. The number of dimensions did not affect their preference for more or fewer candidates.

The attitude most affected by information load was the preference for fewer or more dimensions. Interestingly, this measure was more affected by the number of candidates than by the number of dimensions. After working with 3 candidates, the assessors wanted more dimensions. After working with 12 candidates, they asked for significantly fewer dimensions. Varying the number of dimensions, however, had little effect on the desire for more or fewer dimensions.

Conclusions and Implications

1. Increasing the number of candidates decreased the accuracy of group decisions in a personnel-selection task.

2. Assessors self-reports of confidence and confusion did not reflect the decrease in accuracy produced by increasing the number of candidates.

3. The attitudes affected by increasing the number of candidates, were satisfaction, the desire for fewer candidates, and the desire for fewer dimensions.

4. The implication for personnel selection is that assessors can better cope with an increase in information per candidate than with an increase in the number of candidates.

References

Jacoby, J., Speller, D.E., & Kohn, C. (1974). Brand choice behavior as a function of information load. Journal of Marketing Research, 11, 63-69.

Jacoby, J., Speller, D.E., & Berning, C.K. (1974). Brand choice behavior as a function of information load: A replication and extension. Journal of Consumer Research, 1, 33-42.

Sackett, P.R., & Wilson, M.A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. Journal of Applied Psychology, 67, 10-17.

Streufert, S., Suedfeld, P., & Driver, M.J. (1965). Conceptual structure, information search, and information utilization. Journal of Personality and Social Psychology, 2, 736-740.

Streufert, S., & Schroder, H.M., (1965). Conceptual structure, environmental complexity and task performance. Journal of Experimental Research in Personality, 1, 132-137.

Streufert, S. & Driver, M.J. (1965). Conceptual structure, information load and perceptual complexity. Psychonomic Science, 3, 249-250.

Streufert, S., Clardy, M.A., Driver, M.J., Karlins, M., Schroder, H.M., & Suedfeld, P. (1965). A tactical game for the analysis of complex decision making in individuals and groups. Psychological Reports, 17, 723-729.

ASSESSMENT CENTERS & BIO-DATA: COMPETING & COMPLIMENTARY APPROACHES TO MANAGERIAL SSELECTION

CRAIG J. RUSSELL
RUTGERS UNIVERSITY

An innovative method of biodata item generation was used in conjunction with an assessment center to select candidates for retail store management positions. The method of biodata item generation proved to be a highly effective and cost efficient method for generating a valid biographical information questionnaire. Assessment center dimensions were used as stimuli for life history essays from 150 current store managers. Specifically, the managers were given the definitions of three assessment center dimensions and asked to write a brief essay about some prior life experience that they felt was an example of this aspect of management. Essays were content analyzed for critical incidents using the Campbell, Dunnette, Lawler, and Weick model of managerial performance (i.e., antecedent processes -> behaviors -> task outcomes; all of which are influenced by environmental characteristics). The assessment center and biodata instrument were then evaluated in a concurrent validity design on a sample of 748 current store managers. Performance appraisal ratings were made by immediate superiors on performance dimensions of Personnel Responsibilities, Resource Management Responsibilities, Customer Interaction Responsibilities, and Operation Management Responsibilities. Validities range from .28 to .54 with the performance measures. The use of the <u>self administered and scored</u> biographical information questionnaire for assessment center pre-screening to obtain substantial cost savings is discussed. The use of biodata instruments for needs assessment in career development is also discussed.

Development and Initial Validation of a Biodata Inventory in a Merit System Context[1]

Jay A. Gandy and David A. Dye
U.S. Office of Personnel Management

This paper discusses the development and concurrent validation of an empirically scored biographical questionnaire for potential use in selection for entry-level prefessional and administrative occupations. Although "biodata" instruments have long been used in the private sector because of their frequently high validity and low adverse impact on minorities, their use in the public sector has been extremely limited. Reasons for limited use of biodata measures in both the public and private sectors have included concerns about privacy, content and construct validity, stability of empirical validities over time, and occupationally and situationally specific validities (McKillip & Clark, 1974; Owens, 1976; Mabe & West, 1982; Reilly & Chao, 1982; Mumford & Owens, 1987). Concerns about accuracy and faking have also been an inhibiting factor (cf. Mosel & Cozan, 1952; Goldstein, 1971; and Pannone, 1987).

In 1987, we took a fresh look at the issues surrounding the use of biodata and hypothesized that the problems were surmountable. Items could be developed with attention to content and job relevance and screened to avoid invasion of privacy. Stability of validity should be increased by use of a large sample for item keying and validation. Both stability and generalizability of validity should be enhanced by basing item keying on broadly based samples that included virtually all professional and administrative occupations, agencies, and locations; and by using a broadly applicable measure of job performance.

Method

Inventory Development. Development began with a review of the taxonomies of past behavior items compiled by England (1971) and Glennon, Albright, and Owens (c. 1961). To be acceptable items had to deal with events under individual control; have potential relevance to job performance; be verifiable in principle; avoid invasion of personal privacy; and avoid stereotyping by race, sex, or national origin. Although this screening process eliminated most categories of background data, remaining areas were school and educational experience, work history, skills, and interpersonal relations; a pool of 148 multiple choice items were developed in these areas.

Criterion. A supervisory performance appraisal score was used based on the USES Descriptive Rating Scale (DRS). The DRS has a history of successful use and reflects a "generic" appraisal believed to be generally applicable to all occupations covered. The appraisal included multiple choice descriptions

[1] Alice N. Outerbridge and James C. Sharf contributed substantially to this research. Opinions expressed in this paper are the authors' and do not necessarily represent the official policy of the U.S. Office of Personnel Management.

relating to quantity and quality of work, accuracy, job knowledge, and efficiency, plus a summary appraisal. The criterion score was computed as the average across all scales.

Research Sample. Research materials were prepared for distribution to all external hires at the GS 5/7 level for PAC occupations through all appointment authorities for calendar years 1983-1986 who were still on board (13,000) plus a weighted random sample of 2,300 inservice placements based on population size by occupation. Supervisor-employee matches were obtained for 6,300 employees. This sample was found to be reasonably representative of the target population with respect to gender, race and national origin, occupations, and agencies.

Procedure. Research materials were delivered to sampled employees and supervisors nationwide through approximately 900 servicing personnel offices of 35 agencies. Supervisors were requested to allow time on the job for completion of the questionnaire. All participants were assured of confidentiality, and completed response sheets and questionnaires were mailed directly to OPM in preaddressed envelopes.

Validation Method. A double cross-validation design was used entailing the following steps: (1) splitting the total sample into two random halves; (2) developing a scoring key on each half; (3) applying the key developed on one half to the other half; (4) correlating the scores so derived with job performance; and (5) evaluating the degree of "shrinkage" in correlations when a key is applied to an independent sample. The final key was developed on the total sample.

Item Keying and Scoring. Scoring keys were developed empirically based on point-biserial correlations between each item response and the criterion. Empirical keying has been found to be equal or superior to other methods for purposes of personnel selection (Mumford & Owens, 1987). A statistical significance level of .001, and unit weights (0, 1, 2), were used in keying response choices. Additionally, rational decision rules were developed to avoid illogical keying in situations of low response rates to extreme response choices. Keying was conducted independently by two psychologists. Differences were due almost exclusively to clerical error, and all were readily resolved.

Subgroup Analyses. Separate validity analyses were conducted by gender and race/national origin subgroups which had N's of 200 or larger. Fairness analyses were carried out testing for differences in standard errors, slopes, and intercepts (Gulliksen & Wilks, 1950).

Validity Analyses Across Occupations and Agency Settings. Separate validity coefficients were computed for each of 105 occupations. A meta-analysis of this distribution of validity coefficients was conducted applying procedures outlined by Hunter, Schmidt, & Jackson (1982). The conservative "bare bones" procedures were used in which no corrections to validity coefficients are made and only the effect of sampling error on variance in validities is taken into account. Thus, the meta-analysis consisted of computing the mean validity, weighted by sample size; the standard deviation (and variance) of the validity distribution; the standard deviation (and variance) expected statistically due to sampling error; the remaining variance after subtracting sampling error; and the lower limit of the 90% confidence interval.

Similar procedures were applied to evaluate variations in validity coefficients across agency settings. In this meta-analysis the distribution consisted of validity coefficients computed on research participants in each of 28 federal agencies.

Results and Discussion

Score and Criterion Distributions. The score distribution was approximately normal. The criterion distribution was negatively skewed as is typical of performance appraisals but did show greater variance than operational (administrative) appraisals.

Cross-Validation. As described above, scoring keys were developed independently on random halves of the total sample, designated _A_ and _B_, and on the total sample. Scores obtained from these three keys were then correlated with job performance ratings from each of the three groups, resulting in nine correlations. The critical correlations obtained were the independent cross-correlations, as follows: (a) With key _A_ applied to half _B_, _r_ = .33. (b) With key _B_ applied to half _A_, _r_ = .32. For the total sample, the correlation of the total key with the total sample was .33. The correlation of the total key with each half was also .33. Subsequent analyses were based on the 84 items and item response weights of the total key.

Little shrinkage occurred in validity coefficients in the cross-comparisons. Key _A_ dropped from _r_ = .34 to .33; key _B_ decreased from _r_ = .34 to .32. The results provide strong support for the robustness of scoring keys developed on large samples.

Fairness Analyses. Analyses indicated that the biodata instrument is fair both to minorities and gender groups. Comparisons showed insignificant differences between males (_n_ = 3,535) and females (_n_ = 2,757) on standard errors and on regression slopes and intercepts. Small but significant differences in intercepts were found between blacks (_n_ = 916 and whites (_n_ = 4,842) and between Hispanics (_n_ = 310) and whites. These differences indicated a small degree of overprediction for blacks and Hispanics.

In an effort to reduce group differences, 20 items were removed based on item response statistics reflecting relatively low validity for minorities in conjunction with relatively high minority response rates to low-weighted item alternatives. This led to reductions in subgroup differences as follows: The black-white effect size (difference in means divided by total sample standard deviation) decreased from .34 SD to .28 SD; and the white-Hispanic effect size decreased from .19 SD to .09 SD. Male-female differences also narrowed from .25 SD to .11 SD. These rather substantial reductions in subgroup differences were achieved at the cost of a small change in total group validity from .33 to .32.

1[illegible]

Meta-analysis Across Occupational Series.. The mean validity, weighted by sample size across the 105 occupations, was .30. Sample sizes ranged from two to 845, and total *N* = 6,295. The observed standard deviation of the validities was .133, and the standard deviation expected on the basis of sampling error was .113. The ratio of the squares of these values indicated that nearly three-quarters of the variance in the distribution of validities is accounted for by sampling error and that variations, if any, in true validity are likely to be small. The lower limit of a 90% confidence interval for the remaining variance was .21, indicating a very strong likelihood that the biodata instrument is valid for all covered occupations. Given that a single validity coefficient computed across all incumbents, at .32, is higher than the average validity of occupationally specific validities, at .30, it appears that nothing is to be gained in terms of predictive accuracy by treating the occupations separately.

Meta-analysis Across Agencies. Similar results were obtained across agencies. Meta-analysis of *r*'s computed across 28 agency settings (*n*'s ranging from two to 1,288) again yielded a sample-size-weighted mean validity of .30. Here, however, *all* of the variance was accounted for by sampling error.

These findings strongly support the use of a common scoring key across occupations and indicate that differences in agency settings are unlikely to affect biodata validity.

Conclusions

Findings from this research suggest that biographical data can be an effective means for selection within a merit system environment. By effectively implementing a number of controls in the design, development, and keying of biodata inventories, this study found that: 1) levels of empirical validity can be maintained, 2) fairness both to minorities and gender groups can be achieved, and 3) generalizability of validity can be ensured across different occupations and job settings.

Further research is underway to examine other important issues. Since faking of responses has been a perennial concern with biodata, steps are being taken to address and deal with this issue. Other studies are being done to examine the role of underlying constructs.

References

England, G. W. (1971). *Development and use of weighted application blanks* (Rev. ed., No. 55). Minneapolis, MN: University of Minnesota, Industrial Relations Center.

Glennon, J. R. , Albright, L. E., & Owens, (1961). *A catalog of life history items*. Washington, DC: Scientific Affairs Committee, American Psychological Association, Division 14.

Goldstein, I. L. (1971). The application blank: How honest are the responses? *Journal of Applied Psychology*, *55*, 491-492.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika*, *15*, 91-114.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills: Sage Publications, Inc.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, *67*, 280-296.

McKillip, R. H., & Clark, C. (1974). *Biographical data and job performance* (TM 74-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Mosel, J. N., & Cozan, C. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology*, *36*, 365-369.

Mumford, M. D., & Owens, W. A. (1987). Methodology Review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement*, *11*, 1-31.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609-644). Chicago: Rand McNally.

Pannone, R. (1987, September). *The effects of faking on biodata validity coefficients*. Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY.

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

Video Based Structured Interviews

Carla Swander
Seattle, Washington

This presentation will consist of examples and discussion of innovative video based structured interview materials. One example tests for the ability to work cooperatively as part of a crew (linework). Another example is from the nation's largest retail clothing chain.

Structure of Interviews

Video based structured interviews are designed according to accepted testing principles. This includes job analysis and structured rating criteria. Reliability of the process is due primarily to the most traditional aspect which is structure. However, reliability is enhanced by entirely consistent presentation of interview questions.

Questions are presented on video and candidates give free response answers that are evaluated by raters using structured rating criteria. The interviewing process can be accomplished much more quickly, with little negative effect, by allowing applicants to watch the questions beforehand, making note of their answers. They bring their notes into the interview setting and orally discuss the questions and their answers with the raters.

Advantages Over Traditional Interviews

❑ Questions are presented more vividly. Emotional elements can be portrayed.

❑ Complex questions can be asked quickly and simply. Interviewees always understand the question.

❑ Aspects can be evaluated that are difficult, if not impossible, to assess using traditional methods.

❑ Interviewers can devote their attention to listening and rating, rather than asking questions. Less training of raters is required.

❑ The sophistication of the process presents a good image of the organization. This is particularly advantageous for some of our clients in candidate short markets.

Advantages Over Multiple Choice Tests

❑ The obvious advantage of video interviewing over video multiple choice testing is that the questions are free response. Applicant tone of voice, facial expressions, hesitations, and other indicators of meaning can be observed and evaluated.

❑ Answers are not suggested, so questions can be more basic.

❑ Developmental costs are lower. The average number of items in a video based interview is 10-12. Multiple choice tests ordinarily have 50-125 items.

Disadvantages

❑ As in any interview process, lower developmental costs are somewhat offset by more expensive administration costs, most notably rater time.

❑ Also, as with any interview process, the fewer number of items restricts variance and reliability.

All in all, video interviews are an excellent selection tool in certain settings. These are settings in which the video is advantageous in presenting questions and in which the candidate population is small enough to permit interviewing rather than other forms of testing.

Video Testing for Correction Officers

Oscar Spurlin Ph.D. and Carla Swander
Seattle, Washington

In cooperation with the State of Oregon, a video-based selection test for correction officers has been developed. This instrument was designed for use in all types of correctional facilities, and is primarily aimed at testing the ability to interact effectively with inmates. The test is multiple choice format, with the questions presented on video. Preliminary findings indicate a positive reductions in turnover and sick leave and significant correlation with supervisory ratings.

Background

In cooperation with the State of Oregon, a video-based selection test for correction officers has been developed. This instrument was designed for use in all types of correctional facilities, and is primarily aimed at testing the ability to interact effectively with inmates. Reported here are results of criterion-related validation in progress in the states of Oregon and Missouri.

Job analysis, conducted with correction officers, supervisory personnel, and inmates, supplies strong evidence to show that good human interaction skills contribute significantly to a safe and secure corrections environment. One example where this is especially true is in low security facilities where the officer/inmate relationship is the only form of inmate control. The importance of psychological dimensions has been emphasized in court directives requiring organizations to evaluate the psychological fitness of those who apply for work in correctional institutions.

The video-based test is designed as a series of critical incidents where correction officer behaviors have high consequence of error. Preliminary findings indicate that the test is predictive of supervisory evaluations, turnover, and absenteeism. No significant score differences by race or sex have been observed.

Test Format

The test is composed of 118 multiple choice questions presented on video. Examples of job content dimensions covered by test questions:

- Officer response to manipulative behavior,
- Officer response to provocation,
- Sensitivity to the human needs of inmates,
- Perceptual skills,
- Recognition of when the officer should refer problems to someone else,
- Consistency and fairness when dealing with inmates regardless of race, offense, group affiliation or personal bias.

Test Development/Validation

Content Validity Evidence - Job analysis includes:

□ **Critical incidents** - from officers, supervisors, and inmates.

□ **Indepth interviews** - with hundreds of officers, supervisors, administrative personnel, and inmates.

□ **Job analysis surveys**- across a wide population of officers representing various types of institutions.

The job analysis study was designed not only to find the complete range of human interaction skills and requirements of the job but also to provide an overview of all aspects of correction officer work. The job analysis technique we employed was principally a variation on critical incident methodology. This approach is the most effective way to understand the complex behaviors, particularly in the area of inmate interactions, that make up effective day to day job performance.

The objective of the job analysis was to have all concepts relating to job performance (whether these concepts involved traits, tasks, attitudes, or whatever) defined in terms of actual behaviors. Since the test was to portray this behavior, this approach provided unambiguous and accurate information. Alternative approaches, such as task analysis, would not have yielded the detail of information needed to construct a video test.

Composition of Job Analysis Panels

In Oregon, 15 half day job analysis sessions were conducted. Each session averaged 12 subject matter experts. Job analysis participants included:

- Male and Female Correction Officers
- Administrative and Supervisory Personnel
- Male and Female Inmates
- Minorities and Nonminorities
- Personnel from minimum through maximum security facilities
- Personnel from male and female institutions

Results

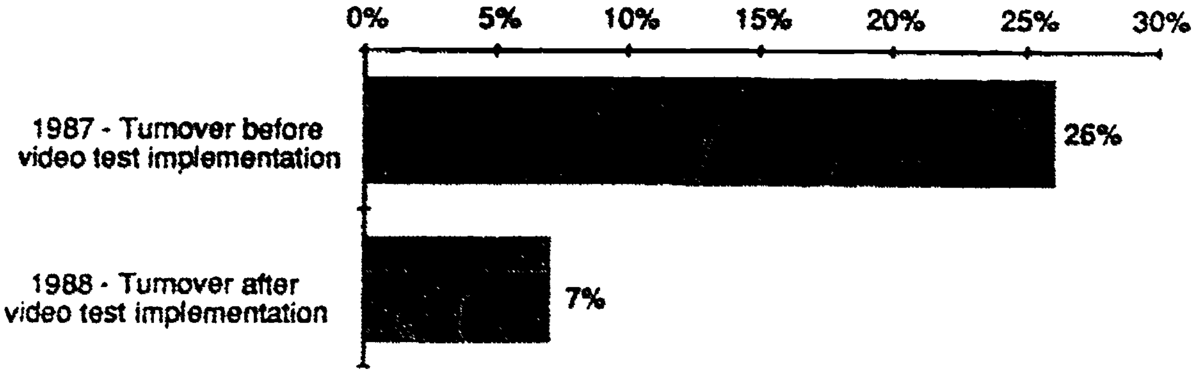
The State of Oregon begin using the Video Test in February of 1988. New hires, selected with the video test through December of 1988, have been compared with a control group of new correction officers hired in the previous year, 1987. Information on reduced turnover and absenteeism is shown in Figures 1 & 2.

Criterion-related validation studies are continuing at various locations. Concurrent validation was undertaken in the State of Missouri. The test was given to 200 current corrections officers. A confidential critical incident based supervisory performance evaluation was used as principal criterion. Data is also being collected on absences and sick leave. Figure 3 demonstrates the relationship between performance on the video test and the overall supervisory rating. The uncorrected validity coefficient is .34.

Figures 4 & 5 illustrate that there is very little score difference by race or sex. The differences displayed here are not statistically significant.

Figure 1

73% Reduction in Turnover Among New Hires

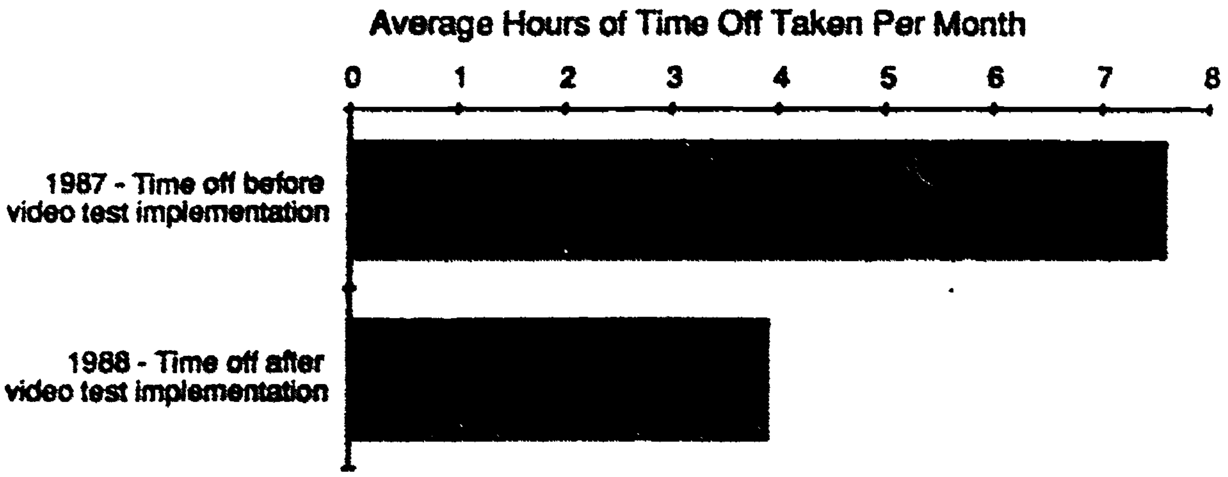


In 1987, prior to video test implementation, of 109 hired, 28 terminated by year end.

In 1988, after video test implementation, of 100 hired, 7 terminated by year end.

Figure 2

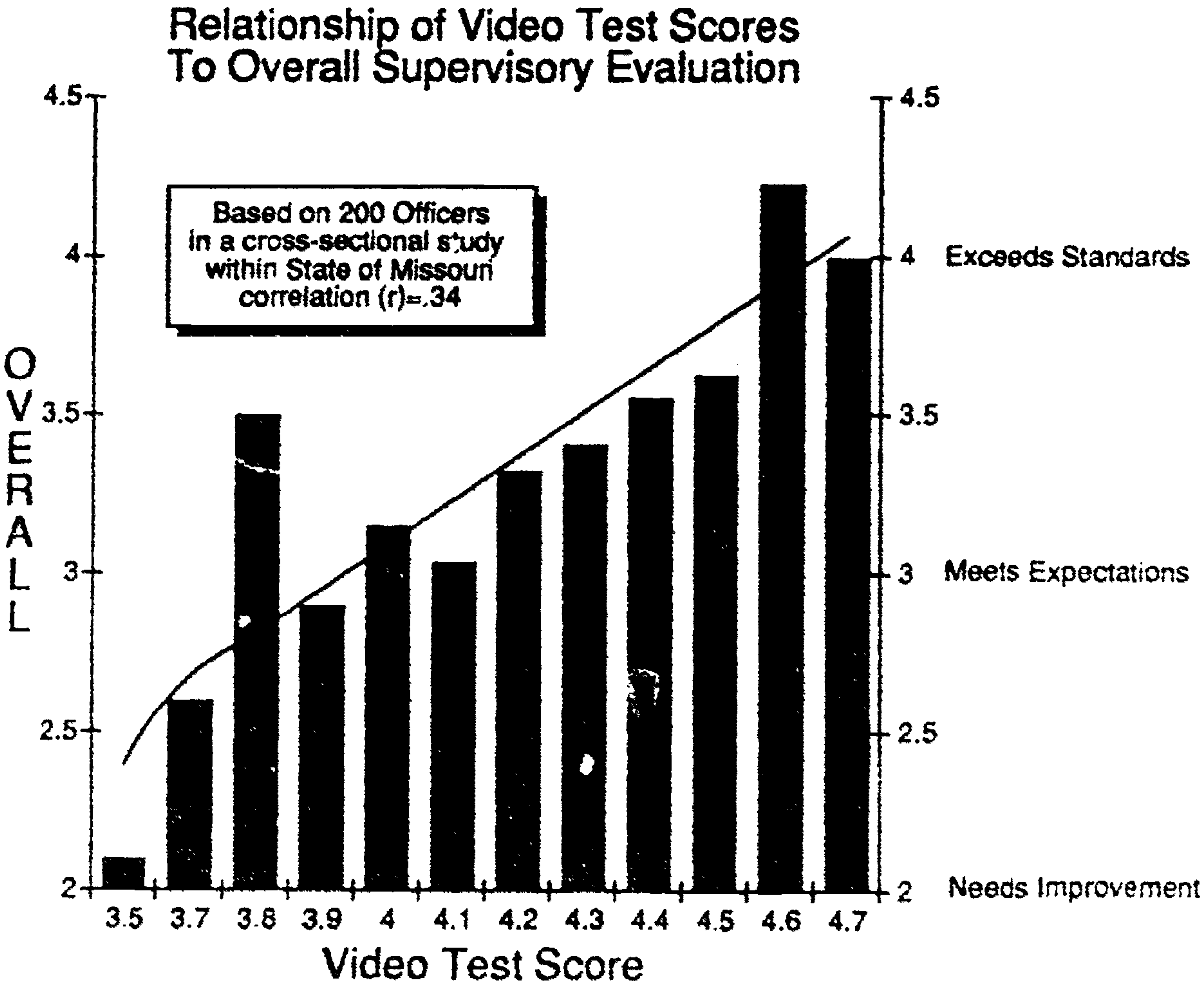
49% Reduction in Total Time Off
Includes Sick Leave and Leave Without Pay



The 109 officers hired in 1987, before video test implementation averaged 7.6 hours of time off per month.

The 100 officers hired in 1988, after video test implementation averaged 3.9 hours of time off per month.

Figure 3



Based on 200 Officers in a cross-sectional study within State of Missouri correlation (r)=.34

Figure 3. The top of each bar represents the mean supervisory evaluation score of test takers in that score interval. The video test score represents the average item score of 118 questions, each with five possible points.

Figure 4

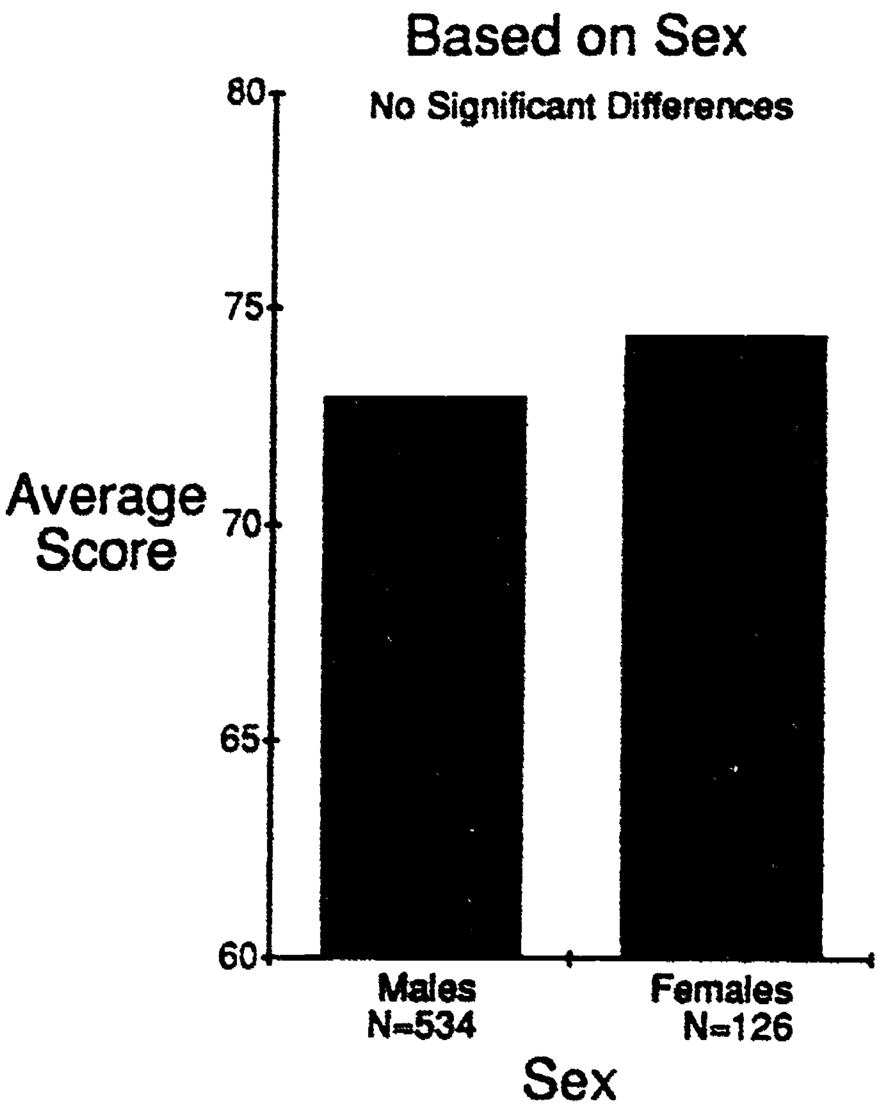
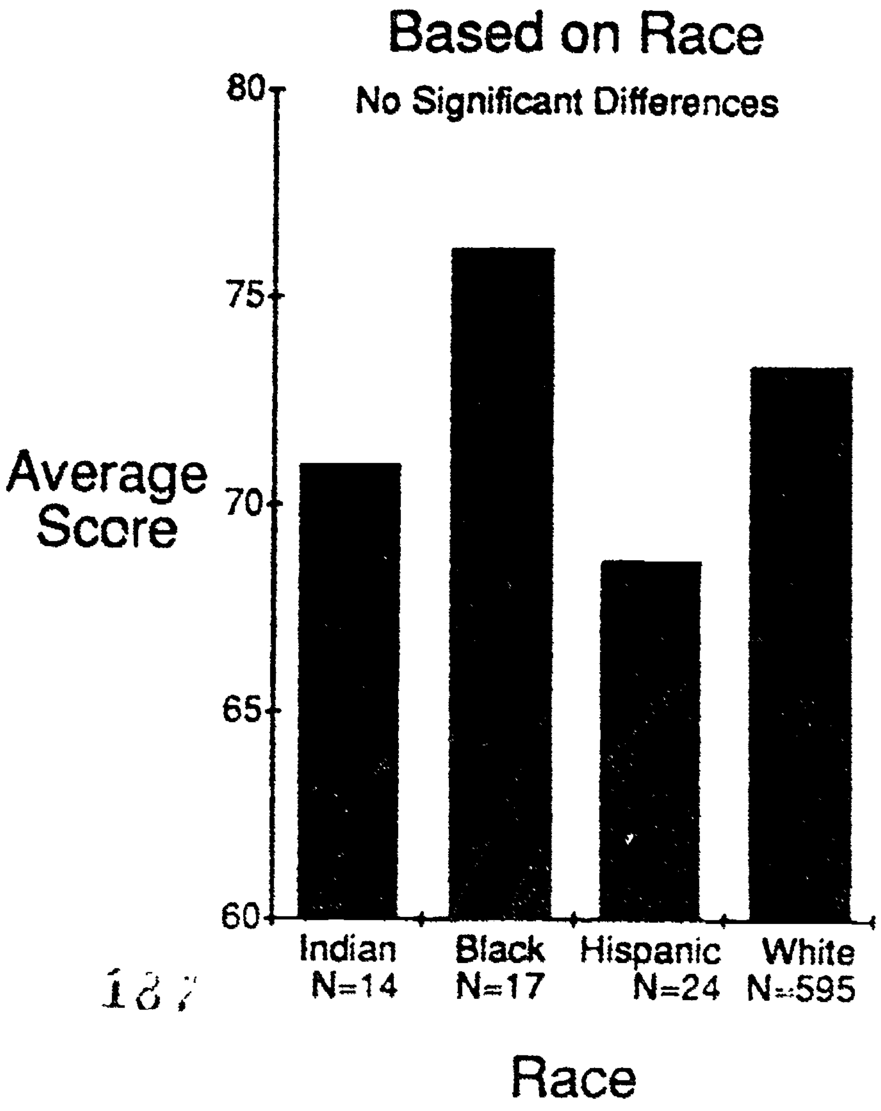


Figure 5



Turnover in the Federal Government:

Implications for Personnel Selection Procedures[1]

Paul van Rijn

U.S. Merit Systems Protection Board
1120 Vermont Avenue, N.W.
Washington, DC 20419

This paper describes the results of an analysis by the U.S. Merit Systems Protection Board (MSPB) of turnover in the Federal Government and discusses the implications of the findings for selection procedures. The analysis was conducted by the Office of Policy and Evaluation, which has statutory responsibility to conduct special studies of the Federal civil service and other merit systems.

There is a growing concern about the ability of the Federal Government to retain valuable employees. This concern is currently being debated in the context of such issues as Federal pay and the image of public service. Turnover is very costly, not only in dollars spent in recruiting replacements but also in lost productivity and errors caused by talent vacuum. There is a growing concern about the ability of the Federal Government to retain valuable employees. This concern is currently being debated in the context of such issues as Federal pay and the image of public service. Some level of turnover, of course, is not only natural but desirable as a way of keeping organizations fresh and vital. Too much turnover or the untimely departure of key personnel, on the other hand, can be debilitating.

Although many Federal agencies have studied selected occupations or have examined turnover within certain departments, there is currently no readily available Governmentwide source of information that can be sifted for answers to shed light on the nature and extent of the problem. The time is right to ask "who is leaving the Federal Government?"

In part, to help answer the question of "who is leaving?" MSPB analyzed turnover for calendar year 1987. The data analyzed were derived from the Central Personnel Data File (CPDF), the automated Federal personnel data system maintained by the U.S. Office of Personnel Management.

[1] Paper presented at the 1989 International Personnel Management Association Assessment Council Conference, Orlando, Florida, June 18-22. The views, opinions, and/or findings contained in this paper are those of the author and do not necessarily reflect the official policies or positions of the U.S. Merit Systems Protection Board.

Turnover was operationally defined as the percentage of employees who left the Federal Government during 1987. Four subcategories of turnover were identified: resignations, voluntary retirements, agency removals, and other separations. The scope of the analysis was limited to the 1.4 million full-time, permanent, white-collar employees whose records are included in the CPDF. This sample represents about 70 percent of the 2 million civilian employees in the CPDF and represents nearly half of all Federal civilian employees. (The records of U.S. Postal Service employees and employees in various national security and other agencies are not required to be reported to the CPDF.)

The results show that 9 percent (119,669) of the full-time, permanent, white-collar work force left the Federal Government in 1987. Over half (58% or 69,298) of these separations were resignations. Only one-fourth (25% or 30,211) were retirements, while agency removals and other separations made up the remainder. The Governmentwide **rate of resignations** was 5 percent, while the **rate of retirements** was 2 percent.

While these average separation rates are relatively low and appear to present little cause for concern, more detailed analyses reveal a more alarming picture. Not unexpectedly, the average ratings effectively masked substantial subgroup differences in the rates of separation.

Consistent with the turnover literature, the highest turnover rates occurred among younger and lower-tenured employees. When turnover rates were analyzed for these subpopulations, the rates of separation were dramatically higher. Total separations during the first year of Federal service occurred at a rate of 25% during 1987, i.e., one out of four new hires left the Government within the first year.

Occupations also vary considerably in rates of turnover. As expected, some large occupations (e.g., nurse, tax examiner, and clerk/typists) had average turnover rates that were almost twice the Governmentwide average. Others, such as computer specialists and engineers, defied conventional wisdom and had average turnover rates that were only about half the Governmentwide average.

By themselves, turnover rates do not fully reveal the magnitude of a retention problem. For example, while low turnover in an occupation may suggest that there is no retention problem in that occupation, there could be serious retention problems, if the few employees who do leave happen to vacate critical positions or are exceptionally high-performing employees. Attrition of these types of employees from these types of positions becomes particularly acute, if there are no suitable replacements in the applicant labor pool--that is, the severity of the attrition problem depends, in part, on the magnitude of the recruiting problem.

To obtain some insight into the quality of the employees who left, we examined their most recent performance appraisal data. Almost 16,000 employees with "outstanding" performance ratings left the Government in 1987--half by resignation. One out of five of the employees who resigned had an "outstanding rating." Although the percentage of outstanding employees who left is only about 1 percent of the Federal work force, the sheer number of these separated employees is not trivial.

While the overall turnover rates tend to be relatively low, the data analysis suggests that there are some identifiable work force subgroups for which turnover rates are high and undoubtedly costly. Certainly, the high overall rate of attrition during the first year of service is cause for concern. Second, some occupations have much greater turnover than others. Thir' the magnitude of the turnover problem must be tempered by the recruiting problem.

The challenges for personnel selection specialists are considerable. First, it is necessary to identify to what extent turnover rates are related to selection practices. For example, is it possible that certain selection practices in certain occupations are contributing to high rates of turnover? If so, can these selection practices be altered to increase employee tenure. The utility of a selection procedure is severely eroded, if new hires do not remain in the organization. Who is more valuable to an organization: an "A+" applicant who remains with the organization 6 months or a "B-" applicant who remains for 10 years? Or do we, perhaps, need a mixture of A+ and B- employees? If so, in what relative proportions do we want to employ these applicants?

To help increase employee tenure, personnel specialists may want to consider including tenure among the job criterion measures in future validity studies. A more complete assessment of the total applicant (of noncognitive as well as cognitive dimensions) may be helpful in selecting employees who will be productive and who will remain with the organization for a useful duration. The inclusion of biodata or other less traditional assessment procedures in the total selection process may be useful in reducing the turnover of high-performing employees.

Clearly, selections procedures alone cannot be expected to remedy all retention problems. It is becoming increasingly apparent that improved recruiting, more realistic job previews, improved selections procedures, more satisfactory compensation levels, job restructuring, training (of employees and their supervisors), and management practices all play roles in the likelihood that a productive employee will remain with the Federal Government--or any organization. The precise mixture with which these various human resource management tools need to be applied will vary from organization to organization, from occupation to occupation, and from location to location.

It must be recognized that many factors contributing to turnover are beyond the control of the human resource manager, e.g., the general labor market, health problems, or personal problems and interests of the employee. Nevertheless, given the high cost of turnover, the benefits to be derived from a more systematic investigation and management of turnover may well be worth the costs--particularly in light of predictions of increased competition for a decreasing pool of qualified job applicants.

Validation of the Peres and Garcia Technique for Predicting Performance with Letters of Recommendation

Michael G. Aamodt, Devon A. Bryan, and Alan J. Whitcomb
Radford University

Even though letters of recommendation or some form of reference checking are used by over 80% of organizations in the United States (Muchinsky, 1979), research investigating the validity of such techniques has not yielded promising results. In a study of references used in industry, Mosel and Goheen (1959) found that the validity of references was only .13. These results were supported by Browning (1968) who found the validity of references also to be .13 in predicting teaching success.

Research has identified several potential reasons for this low validity. As with the employment interview, factors other than the relevant content of the letters are used to form impressions of the applicant. For example, Cowan and Kasen (1984) found that letters referring to applicants by their first name were perceived as being more positive than letters referring to applicants by a title such as "Mr. Jones" and Knouse (1983) found that letters of recommendation containing specific examples were evaluated more positively than letters without examples.

In an attempt to focus the attention of letter readers on the important content of the letter, Peres and Garcia (1964) developed a technique in which the traits contained in a letter of recommendat'on are highlighted and placed into one of five categories which were developed based on a content analysis of 625 letters of recommendation written for engineering applicants. These five categories and representative traits for each category are:

Mental Agility: Adaptable, analytical, bright, intelligent, logical, resourceful, wise

Cooperation-Consideration: Altruistic, congenial, friendly, helpful, sincere, stable

Dependability-Reliability: Alert, critical, dependable, methodical, prompt

Urbanity: Assured, chatty, cultured, forward, gregarious, sociable, talkative

Vigor: Active, eager, energetic, enthusiastic, independent, industrious, vigorous

As promising as this technique sounds, Peres and Garcia (1964) unfortunately did not attempt to validate the technique. Thus, it is the purpose of this study to investigate the reliability and validity of the technique using two separate samples.

Method

Subjects

The subjects for the first sample consisted of 78 (39 male, 39 female) former graduate students who had completed the graduate program in psychology at Radford University. The subjects for the second sample consisted of 26 (11 male, 15 female) Graduate Teaching Fellows (GTFs) at Radford University. Each GTF was working toward a Master's Degree in psychology and had complete responsibility for teaching two sections of introductory psychology each semester.

Procedure

Graduate Student Sample - Because Radford University uses a reference rating form and actual letters of recommendation are not required, the files of over 200 former graduate students were first examined to locate those students who had at least one letter of recommendation. This process resulted in 78 students for whom at least one letter of recommendation as well as an overall graduate G.P.A. were available. Two of the authors then independently:

1) Read each letter
2) Highlighted the traits in each letter
3) Used the list composed by Peres and Garcia (1964) to place each highlighted trait into one of the five categories
4) When two letters were available, the number of traits in each of the five categories were averaged across the two letters
5) To control for effects of letter length and number of traits used, the number of traits in each category was divided by the total number of traits across the five categories.

Graduate Teaching Fellow Sample - The letters of recommendation for each of the 26 GTFs were analyzed according to the procedure listed above. However, student ratings of the GTF's teaching served as the criterion rather than the GPAs used with the graduate student sample. The rating used was the final question on the rating form already used by the university which asked for an overall rating of the instructor's performance based on a five point rating scale with a "1" indicating poor performance and a "5" indicating excellent performance.

Results

Reliability of Letter Writers

To determine the extent to which letter writers referred to applicants with similar traits, the number of traits in each of the five categories used by each letter writer were correlated. As shown in Table 1, the coefficients across the two samples were fairly low, and in some cases the coefficients were negative. Thus, it would appear that two people writing letters for the same individual will not say the same things.

	Sample	
Trait Category	Graduate Students	Teaching Fellows
Mental Agility	.12	.18
Vigor	-.03	-.58*
Dependability-Reliability	.04	.48*
Urbanity	.08	.31
Cooperation-Consideration	.15	-.17

Reliability of Letter Readers

To determine the extent to which personnel professionals reading each letter agree about the traits that are present as well as the category in which each trait belongs, the number of traits placed by the two raters in the five categories for each letter were correlated. As shown in Table 2, the coefficients were reasonably high for the first sample with the exception of the urbanity category. One of the problems encountered in the first sample was that many of the traits listed in the letters were not contained in Peres and Garcia's lists. Prior to collecting data for the second sample, these new traits were added to the lists and as can be seen from Table 2, the agreement levels increased for four of th^e five categories.

	Sample	
Trait Category	Graduate Students	Teaching Fellows
Mental Agility	.77	.91
Vigor	.86	.64
Dependability-Reliability	.70	.86
Urbanity	.53	.86
Cooperation-Consideration	.87	.96

Validity of the Trait Categories

To determine the validity of the trait categories, the number of traits in each of the five categories for the first sample was correlated with the student's graduate GPA and the number of traits in each of the five categories for the second sample was correlated with the GTF's overall student teaching ratings. As shown in Table 3, the number of traits in the mental agility category significantly correlated with graduate GPAs while the number of traits in the urbanity category positively correlated and the number of traits in the mentalagility category negatively correlated with teaching ratings.

	Sample	
Trait Category	Graduate Students	Teaching Fellows
Mental Agility	.32*	-.44*
Vigor	-.08	.27
Dependability-Reliability	-.13	-.22
Urbanity	.03	.38*
Cooperation-Consideration	.04	.23

Discussion

Our findings indicate that the technique developed by Peres and Garcia (1964) shows promise as a predictor of performance. With both samples, the significant validity coefficients were more than twice the magnitude of the .13 previously found with references.

While the Peres and Garcia categories were successful in predicting both criteria, it is possible that traits might be better classified into a different system. For example, Peres and Garcia's dependability-reliability category appears to be two separate categories; one involving dependability consisting of items such as "responsible" and "dependable" and another involving assertiveness consisting of traits such as "tenacious", "confident", and "determined". This would be an excellent topic for future research.

An interesting finding in our study was that the traits used by two letter writers to describe the same person were not highly correlated. This finding certainly makes sense if one assumes that each letter writer probably observed different aspects of the applicant's behavior as would be the case if one of the letter writers were a professor and the other an employer.

However, even though the low agreement of letter writers is understandable, it does pose potential problems for the validity of the Peres and Garcia technique. Thus, it is important in the future to investigate issues such as the sources from which letters should be obtained as well as the optimal number of letters that should be written for each applicant. Because highlighting and categorizing traits can be a time consuming process, future research might also focus on the development and validation of a trait based checklist such as the one created by Carroll and Nash (1972).

REFERENCES

Browning, R.C. (1968). Validity of reference ratings from previous employers. *Personnel Psychology, 21*, 389-393.

Carroll, S.J., & Nash, A.N. (1972). Effectiveness of a forced-choice reference check. *Personnel Administration, 2*, 42-46.

Cowan, G., & Kasen, J.J. (1984). Form of reference: Sex differences in letters of recommendation. *Journal of Personality and Social Psychology, 46*, 636-645.

Knouse, S.B. (1983). The letter of recommendation: Specificity and favorability of information. *Personnel Psychology, 36*, 331-341.

Mosel, J.N., & Goheen, H.W. (1959). Validity of the employment recommendation questionnaire II: Comparison with field investigation. *Personnel Psychology, 12*, 297-301.

Muchinsky, P.M. (1979). The use of reference reports in personnel selection: A review and evaluation. *Journal of Occupational Psychology, 52*, 287-297.

Peres, S.H., & Garcia, J.R. (1962). Validity and dimensions of descriptive adjectives used in reference letters. *Personnel Psychology, 15*, 279 286.

The Psychology of Managerial Incompetence

by

Robert Hogan

Despite the so-called "cognitive revolution," psychology--especially in the macro areas of clinical, personality-social, and industrial-organizational studies--remains under the influence of behaviorism. This is fine but users of modern psychological theories should be aware of these pervasive behaviorist influences. This is particularly evident in Industrial- Organizational Psychology. Consider, for example, the lessons of the "One Minute Manager" -- these lessons amount to a short course in behavior modification (with the effective manager becoming the master behavior modifier). Consider for example, modern approaches to performance appraisal which are targeted at specific and observable behaviors (rather than the underlying attitudes that give rise to these behaviors). And consider theories of management or leadership that argue, in essence, that there is no such thing as overall managerial talent, that leadership is a function of the situation in which a potential leader finds him or herself located.

In contrast with the situationism of behaviorism, I would like to suggest that the pursuit and achievement of leadership positions in organizations is a function of the actor's personality so that the same kinds of people will be found over and over in management positions in all kinds of different organizations. Specifically, I want to argue that there are some stable and dependable personality correlates of status in organizations--and that is bad empirical news for the situational theorists.

I want to make a second, and even more radical claim, however. I would like to suggest that within the population of managers in modern America, all of whom share characteristics that predict upward mobility, there is a subset of people who are bad for their organizations. And these people share the same negative characteristics independent of the organization in which they are located. Although this is bad news for the organization, it is also bad news for situational theorists whose logic compels the view that incompetence is also a function of the situation, which means bad managers are <u>not</u> responsible for their performance.

There is no one best way to define managerial competence. Possibly the worst way is to define it in terms of a single supervisor's ratings obtained at a single point in time. A better way to do it would be in terms of several supervisor's ratings across a number of occasions over many years. An even better way would be in terms of peer ratings ratings by people at the same level. Another useful way to evaluate managerial performance is by asking subordinates to comment on their supervisor's performance. You can probably guess that in general, manager's are largely evaluated (if they are evaluated) by their supervisor's and that the evaluation is usually rather casual and intuitive.

Consequently, the study of managerial effectiveness is largely the study of the correlates of casual and intuitive supervisor's ratings. Nonetheless, and despite the protestations of the behaviorists, those studies converge, and show the same general themes. And what do they show?

One of the best of such studies was conducted at Standard Oil of New Jersey after World War II, where 443 managers were tested with cognitive tests, personality tests, and biodata questionnaires. Managerial performance was a composite score based on ratings over a career. The research uncovered a syndrome called general potential for management and this correlated nicely with performance. The syndrome involved being forceful, dominant, assertive, confident, and actively taking advantage of leadership opportunities. And that is the standard finding.

So for example, Ghiselli studied 4 groups of managers (N=152), nominated for "supervisor ability", with a personality measure. He reports large correlations between his measure and rated performance, and that performance was related to being bright, initiating, self-assured, decisive, achievement-oriented, and unconcerned with job security.

Finally, Jon Bentz, former head of personnel research at Sears, summarized 30 years of managerial research as follows. Supervisor's ratings of managerial performance at Sears was correlated with being persuasive, socially assured, ambitious for leadership positions, energetic, bright, and having "heightened personal concern for status, power, and money."

There are two points that I would like to emphasize. First, are the characteristics I just listed (self-confidence, assertiveness, ambition, and energy) reliably associated with being a highly rated manager? The answer is "yes". Second, is there more to managerial effectiveness than being self-confident, assertive, etc? The answer is, "Mais certainment", as we say in Tulsa. There is something not very inspiring or even interesting about learning that a particular set of characteristics are reliably associated with managerial performance. It is an empirical fact, and that fact is bad news for behaviorists; nonetheless, these seem to be necessary but insufficient characteristics.

What I have in mind when I say that these characteristics are necessary but insufficient is the fact that, within the total population of people who are managers and who, therefore, are energetic, ambitious, persuasive, and so on, there are a surprising number of people who are incompetent managers. Consider the following: survey research since 1965 reveals a consistent and disturbing trend. Beginning with Hertzberg's (1966) research on worker motivation and hygiene factors, in study after study, across organization, occupation, geographical location, and time period, a substantial number of workers express strong dissatisfaction with their supervisors. Typically between 60% and 75% of the workers surveyed (regardless of when or where the survey is conducted, or the occupational group that is studied), report that the worst or the most stressful aspect of their job is their immediate supervisor. Furthermore, virtually every employed adult reports that he or she had to spend considerable time (years) working for an "intolerable" boss.

The following are some examples that just happened to be on my desk as I wrote this (but there is a nearly endless supply of such examples). Newsweek on 25 April, 1988 had a feature article entitled "Stress on the Job". Based on a variety of surveys and research studies, the article argues that: (1) stress in the workplace is widespread; (2) in terms of law suits, worker compensation claims, lost time and work slow downs, this stress will, in the near future, amount to $18 billion dollars a year in costs to American business, industry, and government; and (3) the principal cause of this stress is a "tyrannical boss". The New York Times on June 14, 1988 reports the results of a nation wide Louis Harris poll. The poll concerned the mismatch in the perceptions of workers and their supervisors regarding the supervisors' performance. These were quite interesting; 77%

of office workers said it was very important that they be given a lot of freedom to decide how to do their work, but only 37% of the executives believed workers thought this was important. On another point, 89% of the employees thought that it was important for management to be ethical in dealing "with employees and the community" but only 41% said this was actually true of their present employers.

My consulting experience and my reading the trade journals suggests that 25% of private school heads turn over each year, 25% of college and university presidents turn over each year, and about 30% of hospital CEOs turnover each year. In the latter case, the executive search firms (who profit from the turnover) report that "most of the CEOs who leave their jobs are terminated, or forced to resign, or quit before they are fired."

I would like to propose, based on the foregoing, that the base rate for managerial incompetence in the United States is somewhere between 60% and 75%. Thus 6 or 7 out of every 10 managers have significant shortcomings as managers, shortcomings that suggest that they may ultimately fail on the job.

I am not interested in the precise figure for the base rate of managerial incompetence. I only propose 60-75% as an opening bid. Whatever the actual figure may be, it will be a good bit higher than it needs to be. What I want to do next is offer some suggestions regarding the reasons for this level of managerial ineptitude. One reason comes immediately to mind-- because it is based on personal experience. People are, for the most part, simply thrust into managerial positions without any significant training. The best technical person in the shop or department is promoted to supervisor. Our business schools and military academies provide no significant leadership training. In most cases people find themselves in supervisory jobs with no prior training--at which point they have to fall back on personal experience and the resources of the culture for guidance. This amounts to old John Wayne, Clark Gable, and Clint Eastwood movies. And once on the job they are given little systematic feedback regarding their performance as a supervisor so that the on-the-job learning process is inefficient if not actually non-existent.

Fortunately, we have more to go on in the analysis of managerial incompetence than my personal speculations. Specifically, Jon Bentz has published his observations regarding the causes of managerial failure at Sears, and Michael Lombardo and his associates at the Center for Creative Leadership have published a perceptive study of managerial derailment in another large organization. Their conclusion is that incompetence is a function of the absence of certain qualities and the presence of certain others. After taking certain liberties with their basic findings, I would like to propose five (5) generic problem sources. Two relate to insufficiencies, three relate to the presence of something that shoulu be absent. I will only mention the first two because I want to spend more time on the last three.

The first generic cause of managerial incompetence is an inability to learn from experience. This is an important problem and, as such, we know little about it. The second generic cause is an inability to think strategically, to plan ahead, to prioritize, or to use resources efficiently. Again, we know little about this in a systematic or formal way.

The foregoing insufficiencies concern cognitive variables. The final three sources of managerial incompetence are rooted in the dynamics of personality. I have in mind here three types of people, all of whom have the personality characteristics that typi*y managers and who, therefore, are readily chosen for managerial positions--because they interview well.

The model for the first poisonous character is Kim Philby, who joined British Intelligence in the late 1930's, who had a brilliant career in that organization, who was by the mid-1950's head of the Russian desk and in charge of relations with the CIA, who lived in Washington and had access to the most sensitive part of Anglo-American intelligence as it applied to the Soviet Union, and who was also a Russian spy. He was brilliant, charming, accomplished, effective, and a traitor. What sort of person is this? First, he had all the requisite characteristics of a manager--he was bright, self-confident, persuasive, etc. He was, in fact, an unusually talented manager. He also had a complex philosophical rationale for his treachery--it concerned the war against fascism that so preoccupied intellectuals in the 1930's. He also had, in crude terms, a personality disorder, which I would describe as paranoid/passive aggressive: this includes a complex and screwy world view, deep seated feelings of resentment, a generalized desire for revenge, and great caution about avoiding overt confrontation. What data do I have to support this assertion? My colleague Warren Jones has been working for about 2 years on an inventory of personality disorders, a psychometrically adequate alternative to Millon's test. The inventory contains scales for Paranoid and Passive-Aggressive tendencies, but these are categories that were designed by a committee rather tha~ psychometric entities. Between those two scales is a coherent scale that Jones calls Resentment. He did a study of interpersonal betrayal, asking people if they had ever been betrayed, and if they had ever betrayed another. This produced some astonishing admissions and confessions. The point, however, is that scores on the Resentment scale (which is common to the scales for Paranoid and Passive Aggressive) strongly predicted events of actual betrayal. This is the first type of flawed managers--charming, talented, ingratiating, quietly gathering negative information about you and your associates and feeding this information to people who don't wish you well.

The second and third type of flawed managers are psychological Siamese twins: both are charming, overtly self-confident, and managerial in appearance, both have, beneath the overt poise and social self-confidence, a hollow core, a reservoir of private insecurity. They differ largely in terms of how they deal with the insecurity. My colleague Robert Raskin calls this the two path model of self-esteem regulation.

The first of these--the second type of flawed manager--is called a High Likeability Floater. These people manage self-esteem by conforming very carefully to the demands of their social environment. Lyndon Johnson's notion that you have to go along to get along is their motto. These people are exceedingly pleasant, congenial, cordial, and attractive. They make wonderful colleagues and charming dinner companions. They are supportive and understanding, they facilitate meetings, and they never complain, argue, or criticize. Because they are so well liked, and because they are such loyalists, they rise steadily in organizations. But they accomplish little along the way because they have no agenda, they stand for nothing, they have no point of view, and they rarely take stands on important issues. George Bush may be an example of this--loyal, kindly, affable, a genuinely nice man but who knows what he stands for?

Sooner or later it becomes obvious that these people won't confront non-productive subordinates, that they won't push for an agenda, that they are charming floaters whose subordinates are characterized by high morale and low productivity. Despite their nonperformance, they are exceptionally difficult to fire because they have no enemies; they do have lots of friends, however, who will be angry if you fire poor old George. The presumption will be that you did it for political reasons. The consequences of this is that the arteries of large organizations are clogged by congenial, cautious, conforming mid-level managers.

The final category of managerial incompetence is the Narcissist. Narcissism is a clinical syndrome that includes exhibitionism, feelings of entitlement (which means that one deserves special privileges and one need not follow the rules that apply to ordinary mortals), egocentrism and indifference to the feelings of others (because they are pawns, extensions of one's own sense of agency). Lyndon Johnson seems to have been an example of this type, a person who compensates for his or her hollow core by being assertively self-confident. Our research shows a steady and surprisingly large relationship between measures of narcissism and conventional measures of leadership and managerial potential.

Robert Raskin has developed the best known modern measure of narcissism, the Narcissistic Personality Inventory (NPI). Research with the NPI illuminates the scary features of this syndrome. Consider these four points. First, the following are typical NPI items. In reading them, ask yourself what kind of person would endorse them in the keyed direction: "I have a natural talent for influencing people" (T), "I am a born leader" (T), "I expect a great deal from other people" (T), "I can read people like a book" (T). Second, the Dominance scale of the California Psychological Inventory (CPI) is one of the best, if not the best measure of leadership and managerial potential ever developed. The scale is conceptualized in prosocial terms. Nonetheless, in several samples, the correlation between CPI Dominance and Raskin's NPI is above .70. *Third*, Raskin and Novacek (1988) report that the following MMPI items have the highest correlations in the MMPI item pool with the NPI: "I am entirely self-confident", "I am an important person", and "In a group I would not be embarrassed to be called upon to start a discussion or give an opinion about something I know well". *Fourth*, persons with high scores on the NPI are described by others who know them well as follows: highly energetic, extraverted, self-confident, competitive, achievement oriented, aggressive, exhibitionistic, manipulative, egotistical, and self-seeking. The point is that conventional indicators of managerial potential are heavily confounded by Narcissism, and that point needs to be acknowledged and dealt with.

Five features of the narcissists' cognitive style are noteworthy. First, it is very hard to give them suggestions or recommendations; to accept the suggestions of a subordinate, in particular, is a sign of weakness. In addition, they are so self-confident that they don't believe others have anything worthwhile to contribute in any case. Second, narcissist tend to take more credit for success than is their due. Third, they are biased to avoid taking responsibility for their failures--those they will blame on their subordinates. Fourth, narcissists typically make judgements with great conviction ("often in error but never in doubt"). As a result, they tend to dominate group discussions and to exercise more influence than they should in group meetings. Finally, narcissists are exhibitionistic and self-promoting; they tend regularly to self-nominate and therefore quickly move into any power vacuum.

Narcissists are motivated primarily by a need for recognition and less motivated by a need for achievement. They are particularly good at ingratiating themselves with their superiors; at the same time they exploit their subordinates who are, after all, merely tools to be used by the narcissist in his or her pursuit of recognition.

In dealing with their subordinates, the crucial feature of narcissism is exploitation. Overtly the narcissist may be gracious and civil--to the point, even, of distraction--but the civility is part of the act and has nothing to do with the recipient. On the other hand, the narcissist may be rude and abrasive--especially if he or she is from the northeast. Either way, civil or abusive, the subordinate is an object to be manipulated in the narcissists' pursuit of recognition.

Let me conclude by restating four key points. First, the behaviorist model of management is wrong; there is a set of distinctive characteristics that typify managers regardless of the type of organization. Second, these characteristics are necessary but insufficient precursors of managerial competence. Third, within the group of confident, charming, assertive, and persuasive people who fill the ranks of management are three subtypes, each of whom has a lot of upward mobility, potential to do some damage, and potential ultimately to derail. Finally, these people typically do well in interviews, assessment centers, and other beauty contests that are used to find high level managers. The lesson for assessment specialists is to begin thinking about screening for subtle (not overt) flaws of personality and character located behind the facade of confidence and charm.

Presidential Address:
Employment Testing: A Public Sector Viewpoint*

by Joel P. Wiesen, Ph.D.

At our last conference, WRIPAC sponsored a presentation by Dean Gifford who was then Chair of the National Commission on Testing and Public Policy. That presentation created considerable discussion at the conference. IPMAAC decided to offer to assist the Commission in its information gathering. This fit with the Commission's plan to solicit papers as part of a three year study of the role of testing. We offered to provide two papers and the Commiss.on accepted our offer. Both of these papers will be issued shortly as IPMAAC Monographs. One paper, entitled "Recent Innovations In Public Sector Assessment", was prepared by Charles Sproule. My presentation today is based on the other paper, entitled "Employment Testing: A Public Sector Viewpoint", which I coauthored with our immediate Past President, Nancy Abrams, and our President-elect, Sally Mc Attee.

Legal, Political and Social Factors

Unique legal mandates and public expectations have led the public sector to develop personnel assessment, selection and promotion methods and systems which are unknown in the private sector or greatly modified from private sector practices.

Specific federal, state, and local laws and strong public expectations combine to require extreme levels of openness, fairness and accountability in all aspects of testing and the resulting personnel actions. The effects of these mandates are pervasive, beginning with public announcement of position openings, continuing with competitive testing of all applicants and limits on discretion in appointment and promotion decisions, and ending with appeals of any or all of these steps. These laws and expectations greatly circumscribe public personnel selection practices.

Public announcement of job openings often leads to large numbers of job applicants, placing practical restrictions on the type of examination which may be employed. Further, the number of applicants is not known when the examination is planned, and there may be little room to tailor the assessment process to the number or quality of applicants, except based on past experience.

Another common feature of civil service hiring is the requirement that all applicants be allowed to compete fairly for job openings. Practicality is not a consideration. Thousands of applicants may be tested for a mere handful of job openings. Again, since the form and content of the examination is typically announced in advance, this can result in very costly and time consuming tests with little additional value in terms of delivering government services.

Together with the mandate for merit-based selection, our practices of public announcement of examinations and open competition on examinations explain the widespread reliance of government jurisdictions on the machine-scored, written, multiple-choice test for civil service examinations.

Merit selection is the cornerstone of all civil service systems. Typically the civil service legislation requires that persons who are appointed be shown to be capable of performing the primary and dominant duties of the position, and assessment is restricted to those areas which can be measured reliably and fairly.

Omitted from the testing process are those areas which cannot be reliably and fairly tested, including various personality variables such as honesty, willingness to take risk, motivation, and willingness to assume authority. Personality factcrs are typically not considered amenable to reliable and valid testing, nor thought to show demonstrable and robust relationships with job performance. Personal attitudes and preferences are typically and similarly omitted for two reasons. First, measures of these areas are so open to faking and subjective evaluation that they are thought not to be capable of fair and reliable measurement. Second, these areas may not fit under the legal mandate to test the KSA's required to perform the job.

The standard for appeals varies among jurisdictions, but the orientation of the civil service appeals body is usually very down to earth. Esoteric tests and testing theory are usually foreign to these groups. The most respected personality tests are viewed with suspicion.

The typical civil service appeal process makes for very conservative testing practice. Test methods which rely on very subjective judgment, or with very modest levels of validity evidence, or tests unusual to the jurisdiction are usually shunned in favor of testing methods which will be easier to defend before the civil service appeals body.

Comparison of Selection in the Public and Private Sectors

Personnel procedures in the public sector are often compared to those of the private sector. But let's be clear that we operate with different goals. A civil service agency has a responsibility to the public which goes beyond that seen in the private sector. This responsibility goes beyond calm labor-management relations to fair and open, as well as effective, personnel selection and promotion.

Unlike the private sector, both the public at large and the applicants for public sector jobs expect and demand a selection process which is straightforward, logical, fair and open. Have you ever heard of an unsuccessful private sector applicant demanding a hearing before the company president to see if sound, fair hiring practices were followed? The public sector is legally mandated to adhere to the highest standards in selection and promotion; the private sector does so whenever possible and practical.

A second difference concerns locus of power. There is no one person who heads the governmant; rather there is a planned division of power. At the state level, the legislature mandates and funds the civil service hiring program. The governor then carries out this mandate as he/she sees fit, often embellishing on or ignoring portions of the mandate. The courts have responsibility for the final review of the testing and hiring actions of the governor (and his or her appointees and the employees in the executive branch).

Due to this division of power, it is quite difficult to effect change in the public sector. Even if a jurisdiction wishes merely to pilot test a new approach to personnel selection, it is probably necessary both to suspend some departmental rules and regulations on civil service hiring and to get specific permission from the legislature. Substantive change in civil service law probably entails a full-blown political process, involving public hearing and legislative hearing and action. And in the process of lawmaking, logic and scientific merit are not the only relevant factors; many diverse political and social interests must be recognized. This often makes change very slow, no matter how sensible and practical the change appears to the civil service agency.

Constraints On The Use Of Testing Methods In The Public Sector

Because of the unique perspectives of the public sector, there are special constraints on the feasibility of various types of tests and the use of those tests. Applicants are usually able to question or appeal the testing process and employers must be able to explain why a particular type of test was used and how it was scored. Public sector employers are concerned that applicants accept the testing process. In addition, applicants expect to be treated fairly and equitably and expect that their rights to privacy will be respected by public employers.

Areas of Greatest Need for Psychometric Research

There are two topics which stand out from all others as being in crying need for reevaluation and progress. These are measuring the affective domain, and selecting supervisory and management personnel.

Measurement of the Affective Domain

As mentioned before, tests used in the public sector for personnel selection and promotion generally are cognitive in nature. They test the knowledges, skills and abilities needed to perform various jobs. Rarely do tests attempt to measure the affective domain, which is concerned with personality or emotions. The affective domain might even encompass a person's social or political orientation. There is a great need for research on innovative measurement techniques which both assess the affective domain and meet the special concerns of testing in the public sector.

If we assume that the findings of validity generalization studies are correct, then cognitive tests provide us with valuable information about a candidate's likelihood of success on a job. These studies also indicate that cognitive tests show only part of the picture; a major portion of the factors which lead to job success are not measured by cognitive tests.

In discussions with managers and supervisors concerning what makes for a successful employee and what factors differentiate the most successful from others, the factors cited include both cognitive and personality. Personality traits such as reliability, dependability, conscientiousness, ability to get along with others, ability to work as part of a team, incerest in the work, self-motivation and willingness to work extra hours when necessary are often cited by managers when discussing job success in a wide variety of jobs. We must find ways to measure these areas in a fair, reliable, valid and practical fashion.

There is another reason why tests measuring the affective domain are desirable. Cognitive tests tend to have an adverse impact against some minority group members. Tests measuring the affective domain, to the extent that there is sound research on this, seem to have little adverse impact.

Measurement of the affective domain in the public sector is particularly difficult because employers must justify use and scoring of tests. Tests measuring the affective domain appear to applicants to be scored subjectively, without a clear right or wrong answer, and easily faked. For non-cognitive tests to be usable in the public sector, these concerns need to be addressed. Creative new measurement ideas are needed to fill an important void. If developed and thoroughly researched, measures of the affective domain could address some of the most significant areas of concern for public personnel assessment, namely an increase in validity and a reduction in adverse impact on minority group members. This research is a difficult but necessary undertaking.

Selecting Supervisory and Management Personnel

Many jurisdictions use civil service examinations to help decide whom to promote to supervisory and management positions. This is particularly true for promotion in the police and fire services. The tests used for this purpose range from written tests of supervisory and subject matter knowledges, to oral examinations, to evaluations of training and experience, to practical exercises and assessment centers of various composition. Despite over 50 years of research and development in personnel selection methods, there is a mere handful of criterion-related validity studies of tests to select supervisory and management personnel in the public sector.

The public sector needs a method of selecting and promoting supervisory and managerial employees that is practical, commonsensical, fair, reliable and demonstrably valid, based on criterion-related as well as content validity studies. Current methods assess diverse areas using diverse tests. Job knowledge tests are sometimes criticized as not covering all aspects of the job (i.e., ignoring the application of knowledges which are measured in the abstract). Traditional training and experience evaluations are criticized for giving undue credit to education, or not being able to differentiate between levels or quality of past job performance. Personnel performance evaluation systems are often short-lived and riddled with faults. Assessment centers are an art rather than a science. Methods abound, criterion-related validity studies do not.

Let us look briefly at the dilemma of selection of supervisory and managerial personnel.

Unique factors in selection for supervisory and managerial jobs

There are at least two reasons why promotion or selection for supervisory and managerial jobs is different from selection for other jobs. First is the change in the nature of the skills needed and the related difficulty in not promoting the "best worker". In a typical promotional setting, say from a Junior to a Senior Computer Programmer, the person with the best programming skills is the logical choice for promotion. But for promotion to a supervisory position, it may be that a person other than the best programmer should be chosen. This may not sit well with the Junior Programmers and is one source of difficulty. Second, and more important, is the difficulty we have in describing and measuring the skills which are needed in the supervisory or managerial job.

There is little consensus on the KSAP's which are required to perform supervisory and managerial jobs. This is surely due, in large part, to the multiple ways such jobs can be successfully approached, indicating that there may be more than one set of KSAP's which qualifies one to perform such jobs. It is also partly due to the role leadership plays in these types of jobs and the lack of agreement in the scholarly literature on theories and methods of leadership. For example, one viewpoint suggests, with some empirical support, that policies, rules, employee motivation and professionalism, and other organizational and individual attributes may serve as substitutes for leadership. Further, even if there were agreement as to the determinants of success as a leader in one type of organization, such agreement might not hold for other types of organizations. For example, it may well be that civilian police departments are qualitatively different from the military. For example, unlike the military, the entry-level police officer has the greatest latitude in carrying out orders. And both police organizations and the military may be qualitatively different from an organization like AT&T. It is intuitively likely that KSAP's which underlie success as a leader in AT&T, in a sales organization, in a manufacturing organization, and in a police department do not overlap completely.

The leadership literature is constantly advancing. The Ohio State studies identified two major, independent factors of leadership, consideration and structure; one focused on people as human beings, and the other on job tasks. Recently it has been suggested that the Ohio State research may only be valid for supervisors in manufacturing environments and that the studies were limited in the types of leader behaviors considered. Focus of much research now has shifted to transformational aspects of leadership considering how and under what circumstances a leader changes, rather than satisfies, the motivational characteristics of the employees.

There are a number of well supported theories which deal with aspects of leadership, such as reinforcement theory, equity theory, goal setting and decision making. However, the integration of these theories into one verifiable, comprehensive theory has not yet been accomplished. So the proper approach for a leader to use in any situation is a matter without firm empirical guidance.

Further, there is little agreement on what variables of the leader and the situation affect each other, despite considerable agreement that there is some interaction. This matter is further complicated by the absence of a standard technology for describing differences between supervisory or management job levels and assignments, the differences between organizational structures and organizational leadership styles, and the implications of these differences for personnel selection.

There are few reports of criterion-related validity studies for the selection of public sector supervisors or managers in the general (journal or textbook) literature. I know of three criterion- related validation studies which have been published as technical reports. One such study was conduced by the U.S. O.P.M.

which reported the development of a generic test for the ranking of applicants for trades and labor supervisory jobs. The test covers 31 areas such as:

1. Interest and ability in applying up-to-date job practices
2. Learning and reasoning
3. Flexibility
4. Knowledge of the job as required for a supervisor
5. Checking on work progress
6. Getting information from employees and acting on it
7. Helping employees with personal problems

McCann Associates completed two studies a while ago providing criterion-related validation evidence for two multiple-choice, job knowledge promotional tests for the police and fire services.

Approaches to selecting supervisors and managers

There is a plethora of approaches to selecting people for supervisory and management jobs. One text summarizes validity evidence for measures such as: cognitive ability tests (e.g., verbal reasoning and mathematical ability), objective (written) personality and interest measures, projective techniques, biodata, peer assessment and business games, among other methods. I will consider briefly one of the more popular assessment approaches, the assessment center.

Despite considerable logical appeal and anecdotal evidence, and a few major criterion-related validation studies, there are several reasons why the assessment center method is not yet a solution to the dilemma of selecting supervisors and managers.

First, the assessment center method includes measurement techniques which are too diverse to be validated or otherwise treated as one selection instrument (or perhaps not even as one method). The variation in assessment centers is at least as great as the variation in multiple-choice tests. For example, the classic AT&T assessment center lasted three and one-half days, included at least nine types of measures (paper and pencil tests, in-basket tests, projective personality tests, clinical interviews with psychologists, group problems, leaderless group discussion, a personal history questionnaire, an autobiographical essay and a self description essay), and rated some two dozen areas (e.g., organization and planning, decision making, creativity, human relations skills, personal impact, behavioral flexibility, tolerance of uncertainty, resistance to stress, scholastic aptitude, range of interest). In comparison, an assessment center for promotion to police sergeant or lieutenant may last a few hours and consist of an oral presentation, a leaderless group discussion and an in-basket exercise.

Second, there is very considerable variation in what dimensions or areas are targeted for measurement by a given assessment center. These may range in number from half dozen to a dozen (or even two dozen) areas. The areas measured from assessment center to assessment center also vary widely in name and definition. The areas measured do not seem to be defensible as pure psychological constructs. This is particularly troublesome if we must meet the standards for construct validation which appear in the Uniform Guidelines on Employee Selection Procedures.

Third, to the extent that personality measures of various types, and projective personality tests in particular, are included in an assessment center, the practical nature of the civil service examination mandated for some jurisdictions is compromised. Further, the acceptance by applicants may be problematic. (Imagine answering an appeal by telling an applicant that on the Thematic Apperception Test he/she told a story about failure which was not given as much credit as another person's story about success.) /

Fourth, it is not clear what assessment centers measure. Originally they attempted to measure the requisite KSAP's. However, over the past few years it has been noted that the ratings of different areas within a given assessment exercise are more highly correlated than the several measures of one KSAP derived from different exercises. Thus, the assessment center ratings may be exercise specific rather than reflective of underlying KSAP's.

Fifth, some researchers have suggested that assessment centers do not measure the ability to perform on the job so much as capture the organization's policy for promotion. A further criticism is that in capturing an unfair promotion policy the assessment center may be unfair in its ratings.

Sixth, there is a considerable debate as to whether a final overall consensus is better than a statistical combination of scores from individual exercises. I understand that the recently revised guidelines for assessment centers now reflects, in part, the growing body of literature which shows that statistical prediction is equal to or better than judgmental prediction.

At worst, assessment center methodology is expensive and bewildering in its complexities, and the results are suspect. At best, assessment center methodology is the most fair and valid approach to supervisory and management selection. Unfortunately, with the current level of scientific knowledge and with the current technology, we can not say where the assessment center method falls between these two extremes.

Need for a technology of selection of managers

The public sector has an immediate need for an agreed-upon technology of selection of supervisory and management personnel. The method must meet the usual psychometric/social/legal requirements of reliability, validity, utility, legality/fairness, and ease of use. Of course, it must reflect public policy as determined by the lawmaking bodies of the country, such as lowest possible adverse impact, defensibility under the Guidelines and related state and Federal EEO laws, rules and regulations. It must be capable of secure use, offering no applicant an unfair advantage, either in reality or in perception. It must also be intuitively considered valid by applicants and other interested parties (e.g., news media), and be capable of objective scoring. Finally it must meet the requirement of many civil service jurisdictions that the test be practical in character and deal, insofar as possible, with the actual duties of the position.

The existing literature in this area does not meet this ideal for two reasons: First, leadership and supervision are not fully understood constructs, or sets of constructs, and, second, the field has not yet agreed upon common definitions for the germane areas to measure, nor which tests or types of tests to use to best measure them.

Despite this lack of scientific clarity, personnel assessment professionals in merit systems are continually being asked and even required to develop sound and defensible procedures for promoting or selecting people for supervisory positions and to do so without unnecessary expense.

I would like to see an integration of the literature which results in a virtual technology, or authoritative guidebook, to practice which deals with such topics as:

- the definition of areas to assess, and the degree of and the need (if any) for factorial purity of such areas.
- the relative weight to be given job knowledge, general cognitive ability, supervisory/leadership skills, interpersonal skills and personality variables.
- whether and how the weights given to the various components should vary with the nature of the job.
- which job requirements can or cannot be reasonably expected to be learned on the job (e.g., job knowledge, general cognitive ability, supervisory/leadership skills, interpersonal skills, personality variables).
- the relative merits of situation-specific test questions as compared to general or pure tests of cognitive ability or problem-solving.
- the relative merits of tests of knowledge and understanding of principles and practices of supervision and leadership versus other approaches to measuring these areas, such as written or video presentation of situational questions, or simulation exercises.
- the degree to which people can be expected to learn supervisory and management skills on the job.
- the appropriate deference to be given to the management style of the organization. (For example, should a selection process for an autocratic organization be different from that for a participative organization?)
- the degree to which the nature of the organization affects the KSAP's which are predictive of success on the job. Should the differences between sales, service, manufacturing, and public safety organizations affect the selection of supervisors and managers for these organizations?
- the extent to which grading criteria determined by groups of subject matter experts (SME's) are consistent over time, consistent across groups of SME's, and objectively correct?

There is a strong need for a synthesis of the scientific literature concerning selection of supervisory and management personnel and the development of a formal, published technology for using that body of scientific knowledge as it now exists. Other disciplines with fast changing bodies of knowledge develop technologies for practice. (For example, the Center For Disease Control recommends specific practices for handling infectious biological materials. These in turn are adopted by hospitals across the country.) Whatever the level of scientific knowledge, public sector personnel selection needs to conduct personnel selection today; we need to apply the existing body of knowledge. Nothing less than a technology of testing is needed, based on sound psychometric and psychological theory and research. Unfortunately, funding for such research and development is lacking.

Call for Systematic Funding of Research and Development

In 1970, Congress passed the Intergovernmental Personnel Act (IPA). Under IPA funding, the research components of many State and local personnel testing agencies were begun or grew significantly. With the elimination of IPA funding, these functions have been severely cut back.

With support from IPA, consortia of State and local personnel testing agencies were founded: NEPPC, MAPAC, SERPAC, GLAC, MINSKIE, RESPAC, and WRIPAC. In 1989, only three of the consortia, WRIPAC, MAPAC and GLAC continue to exist.

We urged the Commission to consider calling for a restoration of funding for the Intergovernmental Personnel Act to provide the resources needed to continue advancement in personnel testing in the public sector.

Conclusion

At one point public jobs were bought and sold in pubs and in the halls of legislatures and Congress. Now most government jurisdictions use formal merit systems to select and promote their employees. These merit systems are the epitome of a bureaucracy: run in accordance with numerous and relatively inflexible rules, but treating all concerned openly, equally and fairly, if slowly. Unfortunately, the same rules which promote equal treatment have also become entrenched and difficult to change. At the same time as there is extremely limited funding to promote change, there is a very complex professional and legal literature concerning employee assessment and selection.

To allow these government merit systems to improve involves overcoming four common difficulties:

- poor and uncertain funding for research
- inadequate base of knowledge and technology, particularly regarding testing the affective domain, and designing tests to predict supervisory and managerial job performance
- inadequate training of assessment staff
- laws which constrain change

All of these can be remedied. The first three were improving rapidly until the Congress eliminated the IPA program. We call for a new IPA to continue that work. To begin to address the last area of difficulty, model legislation should be developed by an influential body to demonstrate a reasoned and legitimate degree of flexibility in a merit selection program. Then this model legislation should be advocated by Federal and state agencies.

If Congress passes a new IPA program, progress in personnel selection at the state and local level will be dramatic. Much of the foundation for such change is in place. A new IPA program would provide the mechanism for refining and disseminating these improvements.

* Most of the Presidential Address which does not appear in this summary may be found in a forthcoming IPMAAC Monograph of the same title, which will be mailed to all IPMAAC members.

KSA Based Minimum Qualifications

Alan N. Machtinger
Manager of Applicant Services/Selection
Delaware State Personnel Office

INTRODUCTION

Like most jurisdictions and businesses, the State of Delaware attaches minimum qualifications to specifications describing each job title (i.e., class of work). These minimum qualifications describe minimum entry requirements for education, training, experience, licensure, verified clerical proficiency and, in rare instances, height, weight or age.

Following an extensive review, the State Personnel Office and State agencies recognized a number of inherent and significant problems with traditional minimum qualifications. These problems include a lack of validity and reliability, a tendency to exaggerate minimum requirements to enhance pay grades, the likelihood of unfairly excluding women and minorties and difficulties in defending and explaining job requirements.

As a result, the Office transferred responsibility for minimum qualifications development from the Classification Unit to the Selection Unit. This transfer reinforced the importance of minimum qualifications as an employment test. In addition, minimum qualification decisions would be totally separated (both organizationally and temporally) from paygrade decisions. The Selection Unit was given the mandate to (1) develop a minimum qualifications process which would be valid, reliable, non-discriminatory, defensible and explainable and (2) develop such minimum qualifications for all State merit classes within a two to three year period.

KSA Based Minimum Qualifications

The Selection Unit approached minimum qualification from a classical test development perspective. As a test, minimum qualifications must be both job related (i.e. related to job content and successful performance on the job) and non-discriminatory. By job related, we must be reasonably certain that those individuals who satisfy the minimum qualifications will be able to perform at least at a minimally acceptable level on the job. Conversely, we must be equally certain that those individuals who do not satisfy the minimum qualifications are highly unlikely to be minimally acceptable job performers.

Like any employment test, minimum qualifications mut be based upon a job analysis. This procedure requires a focused job analysis to identify KSA's which satisfy the following conditions:

1. Minimum Qualifications KSA's must be important to job performance. Only those KSA's which are absolutely essential for successful performance should be included.

2. Minimum Qualifications KSA's must be required at entry on the job. KSA's which may be learned or acquired in a reasonable time frame should not be included.

3. Minimum Qualifications KSA's must be discernible from applications and resumes. For example, while interpersonal skills may be important to job performance and needed at entry, there is no way a rater can draw inferences regarding those skills from application information.

In order to satisfy the Minimum Qualifications, applicants must satisfy all of the Minimum Qualifications KSA's. Absence of one KSA is justification for rejection since each KSA is important for job performance and needed at entry.

Each Minimum Qualification should be accompanied by a rating guide which provides additional instructions for agency staff preparing announcements, evaluating applications and conducting selection interviews. Rating Guides may include explanation of individual KSA's, definitions of terms, examples of experiences, training or education that satisfy or do not satisfy particular KSA's, descriptions of examination batteries or structured career ladders, suggestions for announcement language or allowable selective requirements, etc.

Analysts develop KSA based minimum qualifications in the same general manner as they would any other job analysis i.e. reviews of available job material, meetings with subject matter experts, etc. However, in order to develop minimum qualifications quickly and expeditiously, the analyst goes directly to KSA's (no task list) and lists only those KSA's which are needed at entry, essential to job performance and discernible from applications and resumes.

The Selection Unit developed a manual, Knowledge, Skill and Ability Based Minimum Qualifications: A Guide for Developers, to help analysts, agency personnel professionals and subject matter experts understand the system. The manual includes a description of the system itself, examples of KSA based minimum qualifications and rating guides, documentation requirements and discussions of development issues, such as circumstances under which degree and licensure requirements and discussions of development issues, such as circumstances under which degree and licensure requirement are persmissible, incidented or insufficiently related training or experience and documentation requirements.

At the time of this writing, (April 27, 1989), the State of Delaware had completed more than 700 KSA Based minimum qualifications out of a total pool of 1,200 job classes. The system has enjoyed remarkable acceptable from State agencies and applicants. The most frequent complaint has been that we've had to continue announcing some vacancies using traditional minimum qualifications during the transition period.

Research Strategies for the Development of Physical Ability Standards

Osca: Spurlin Ph.D., T.L. Doolittle Ph.D., Barton Daniel M.S.

ERGOMETRICS and Applied Personnel Research Inc.

Abstract

Our research with more than 75 organizations over a ten year period has demonstrated strong empirical support for physical ability testing. We find high correlations with various criteria including injuries, absenteeism, and performance. Discussion includes: ergonomic analysis of physical demands, test validation models, and advantages of standardized exercise tests over work samples. We also present a summary of a number of studies on the effectiveness of screening. We conclude that content-related tests linked to ergonomic principals are the best available option.

□ Physical Ability and Performance

Strength and stamina are factors that are important components of task performance. However, it should be realized that other factors also contribute. Learning the appropriate technique, timing and coordination, flexibility, perception of visual cues, and many other things play a part in successful performance of the tasks. Thus strength and stamina can be shown to be absolutely necessary, but for highly skilled or complex tasks they may not be sufficient. Although we have observed that for most entry level occupations, minimum strength and stamina is much more critical than other physical abilities such as coordination or flexibility.

The approach we are discussing is designed to ergonomically isolate the underlying strength and stamina components of critical job tasks. It might be termed a work sample approach which only samples the work (in a physiological sense) and not the skill.

□ Appropriate Evidence of Validity

Physical job demands, such as caloric expenditure or lifting requirements, can be measured more precisely than most aptitudes or abilities assessed in pre-employment. Therefore, strength and stamina data from job analysis are appropriate criteria for setting standards. In fact, these job analysis criteria are a superior basis for determining tests and minimum standards than more traditional empirical validation. Cut-off scores can be set relative to measurable criteria of critical task performance, rather than on the basis of applicant or incumbent norms, which is a more common but less desirable procedure.

The appropriate research strategy is multidisciplinary. While physical ability test batteries have been developed and validated employing empirical or work sample strategies, a superior approach utilizes expertise in ergonomics and work physiology to define the parameters of safe and effective physical performance.

The research methodology is designed to accomplish two major tasks:

- To demonstrate that strength and stamina are required for successful job performance, and,
- To determine of the acceptable levels of strength and stamina for safe entry into the occupation.

The first task is validation. The second is setting a cut-off. There can be overlap. If the level of strength or stamina required is so low that almost everyone would be expected to perform acceptably, then the test can have little utility or validity. However, there are few industrial or laboring jobs that require no strength or stamina.

Construct-Content

Some researchers have adopted the point of view that validation of physical ability tests needs to proceed on the basis of "construct validation".

Essentially construct validity is used when the characteristic being measured is broad or theoretical and not subject to strict definition, i.e. intelligence or neuroticism. Construct validation requires the accumulation of data indicating that the test is related to a variety of other measures in a way that would support the theoretical premise.

Much of the debate over whether construct validation strategy is necessary for physical ability measures appears to stem from concern over minimum acceptable levels (the cutoffs). The impact of such variables as:

- learning of specific techniques,
- alternative procedures used by men and women,
- on-the-job conditioning.

The above are cited as examples of reasons strength tests may not be representative of actual working requirements.

However, we would point out that these variables can be taken into account during job analysis and usually only affect the amount of strength or stamina required and seldom have enough impact to eliminate these requirements. It is the cutoff score and not the basic necessity of strength and stamina which is being questioned.

We perceive that the concerns which have lead to a debate on appropriate methodology can be addressed by taking an ergonomic approach to studying critical tasks. The necessary strength and stamina components can be effectively isolated from learned job behaviors or skills. Thus, the relation of test content to job content can be directly specified.

Guidelines

Some of our experience in designing physical ability tests might be summarized into these guidelines:

1. The first guideline is that the specific muscle groups employed on the job needed to be the ones tested.

2. The second guideline is that tests should be objective and reliable indices. Candidates should be able to achieve their maximum score without significant training or coaching.

3. The third guideline is that strength requirements should not exceed the limits typically recommended in ergonomics literature. This research recommends 65-75 percent of an individual's maximum strength capacity for occasional efforts and 15-20 percent for repetitious work.

4. The fourth guideline is that it is possible to accurately estimate maximum dynamic strength from repetitions requiring a sub-maximal (and safer), effort.

5. The fifth guideline relates metabolic demand to maximum aerobic power, where it is well recognized that individuals cannot be expected to perform short term efforts at greater than 75-85 percent of maximum or day long efforts at greater than 33-40 percent of capacity, (Astrand and Rodahl, 1986).

6. The sixth guideline is to use standardized exercise tests instead of work sample tests. Some test developers view work sample tests as inherently more valid than standardized exercise tests. We see that standardized tests can be related to key tasks just as accurately. Exercise tests have other striking advantages, including increased reliability, safety, and generalizability.

□ Ergometrics Follow-up Studies

The following results are illustrative of the reported success of pre-employment physical ability screening. We recognize the methodological problems with interpreting this type on data. These are not experimentally controlled comparisons.

Wood Products Industry

Results were gathered for 2 years of entry level hires into laboring positions within lumber mills (principally sawmills and plywood manufacturing). There were approximately 400-500 hires in four states. Plants that had implemented pre-employment physical ability screening on strength and stamina compared the 2 year averages with the year prior to implementation of a physical performance pre-screen. The results are shown in Table 1.

Table 1

Criteria	Tested Group	Untested Group
Turnover		
after 6 months	0%	29%
after 12 months	0%	40.2%
after 24 months	13.5%	50.0%
Accident Rate (Lost time accidents per year)		
After 0-6 months	6.6%	33.6%
After six months		no difference
Throughout the system, supervisors comment that those individuals tested are better fit employees and have fewer accidents.		

Paper Mill

Strength and stamina measures were used at a large paper mill for hiring summer relief workers. Test implementation took place in mid 1985. Comparison of incidents of injury and absenteeism were made with the prior two years summer hires. The results are shown in Table 2.

Table 2

Criteria	Untested Group 1984 N=55	Untested Group 1985 N=32	Tested Group 1985 N=42
Incident rate			
Lost time injuries	82%	34%	0%
Absenteeism			
Days per year	3.8	1.1	1.9

The 1985 group which was untested were rehires from 1984, who were brought back on the basis of supervisory recommendations. The 1985 group would be considered to be superior employees on the basis of supervisory ratings. Thus, the data in Table 2 indicates a very dramatic decrease in the incidence of lost time injuries in the labeled group.

□ Criterion-Related Studies

Manufacturing Plant

We performed a concurrent validity study involving 100 workers in heavy lifting jobs (routine lifting of 40+ pounds) in material handling and construction jobs (Spurlin, Doolittle 1987). This was a large electronics manufacturing facility. In addition to documenting the content-related validity of lifting and aerobic fitness, we obtained medical records on each employee over the previous five years. (There was an in-house medical facility which dealt with all reported injuries). Of principal interest was the total days lost per year for injuries.

For the total sample there was a correlation of -.30 with days lost and lifting strength, and a correlation of -.20 with aerobic capacity. In graphic terms we observed that almost all longer term injuries took place in the group that was below average in terms of strength and stamina.

Wood Products Mills

A second study involved entry-level jobs in the wood products industry (Spurlin, Doolittle 1984). Most jobs were material handling of light to moderate pieces of wood on production lines. The study involved 104 employees in two plants. Tests included a repetitive pulling test and an aerobic fitness test. Criteria included injuries during past 2 years and we also developed a comprehensive supervisory evaluation instrument covering the full range of observable job behaviors.

For the total sample, correlation of strength (pulling) with the overall supervisory evaluation was .31. Correlation with injuries was -.23. These correlations were statistically significant and in the expected direction.

In this setting we also looked at Pre-Post effects of implementing physical ability testing, similar to those case studies mentioned above. Prior to physical ability testing there was a 25% of new hires were lost due to turnovers. The incident rate for lost time injuries was also about 25% (1 of 4 had a recorded injury per year). This rate was higher for new hires, although separate statistics were not available.

A follow-up of 110 employees on the job for six months or more showed:

- There were no lost time injuries of any type, and,
- There had been about 10% attrition.

This type of finding underscores the validity of these tests in very practical terms.

Lineworkers

A third study involved electrical lineworkers (Doolittle, Spurlin, 1988). After thoroughly establishing the relevance of a battery of strength and stamina measures for this very physically demanding occupation, we performed a concurrent validation study with 48 lineworkers. Criteria included confidential supervisory performance evaluations and lost time due to injuries during prior five years.

The multiple correlation of the physical fitness battery with overall performance was .59. Aerobic capacity correlated .19 with this criterion, while the six individual strength tests used correlated significantly between .26 and .41. Lost days due to injury was predicted with a multiple correlation of .46. Aerobic capacity had a significant correlation of -.19, and the six strength tests had a range of coefficients between -.10 and -.29 (three of which were statistically significant).

In the lineworker study we observed that injuries (with greater than one lost day), were almost all associated with employees in the lower half of the strength distribution.

Note that significant findings such as the above and those reported by other researchers are even more indicative when one considers the research limitations resulting from:

- Restriction in range of study group utilizing current workforce.
- Restriction in range of injury criteria (5-10% of workforce account for 90% of accidents).
- Inadequacies of organizational records.
- Variability in job demands and exposures.

□ Conclusion

Our experience with developing physical ability tests over the past ten years has led us to a conclusion (or bias), regarding appropriate methodology.

Having had the opportunity in a number of settings to conduct criterion-related studies, we inescapably must conclude that it is always the ergonomic job analysis which provides the most compelling and useful rationale for developing tests and setting standards. The presence of empirical validity with supervisory ratings or injury data, while gratifying, just does not stack up with empirical evidence of weights lifted, energy cost and other objective data. Psychologists, testing specialists, and others charged with the development of physical ability measurements should not overlook the growing body of relevant literature and expertise in the field of ergonomics.

References

Doolittle, T. L. and Kaiyala, K., 1988. Prediction of maximum dynamic strength from multiple repetitions with a submaximal load. Trends in ergonomics /human factors V . Aghazadeh, F. (Ed.), 767-774.

Logan, G. and Foreman, K., 1961 Strength-endurance continuum. The physical educator, 18, 103-105.

Spurlin, O.L. and Doolittle, T. L. 1984. Development of Job-Related Applicant Screening Procedures. Proceedings of conference of Forest Products Research Society, 7305. p60-67.

Spurlin, O.L. and Doolittle, T. L. 1987. Unpublished Technical Report.

Doolittle, T. L., Spurlin, O.L., and Kaiyala, K. 1988. Physical Demands of Lineworkers. Proceedings of the Human Factors Society -32nd Annual Meeting. p632-636.

The Impact of Physical Standards Projects on Internal Race and Sex Relatiors

Carla Swander and Oscar Spurlin, Ph.D.
Seattle, Washington

Physical standards that are too low place employees, particularly women, at great risk of injury (often permanently disabling). Prejudice toward protected groups increases when standards are too low. This presentation explores implications of physical standards based on numerous studies of race and sex relations, physical demands, and on-the-job injuries.

Physical standards projects are usually well intentioned efforts to address sex, race, safety, and legal considerations. Ensuing politics can be counterproductive to workplace integration and impede good research. These projects must be approached and conducted with great sensitivity for those who are currently on the job.

Physical Capacity and On-the-Job Injury

Workers in the lowest physical abilities groups are a much greater risk of on-the-job injuries. Employees at a 10,000 worker AT&T manufacturing facility were tested for aerobic capacity and lifting strength. These test results were compared with five years of injury data from the facility. The results showed that employees who averaged more than five days injury time loss per year were *all* in the lower half of the physical abilities range. Figure 1 shows the relation of VO2MAX (maximum oxygen uptake, a measure of aerobic capacity) to time loss for injury. Figure 2 shows the relation of lifting strength to injury time loss.

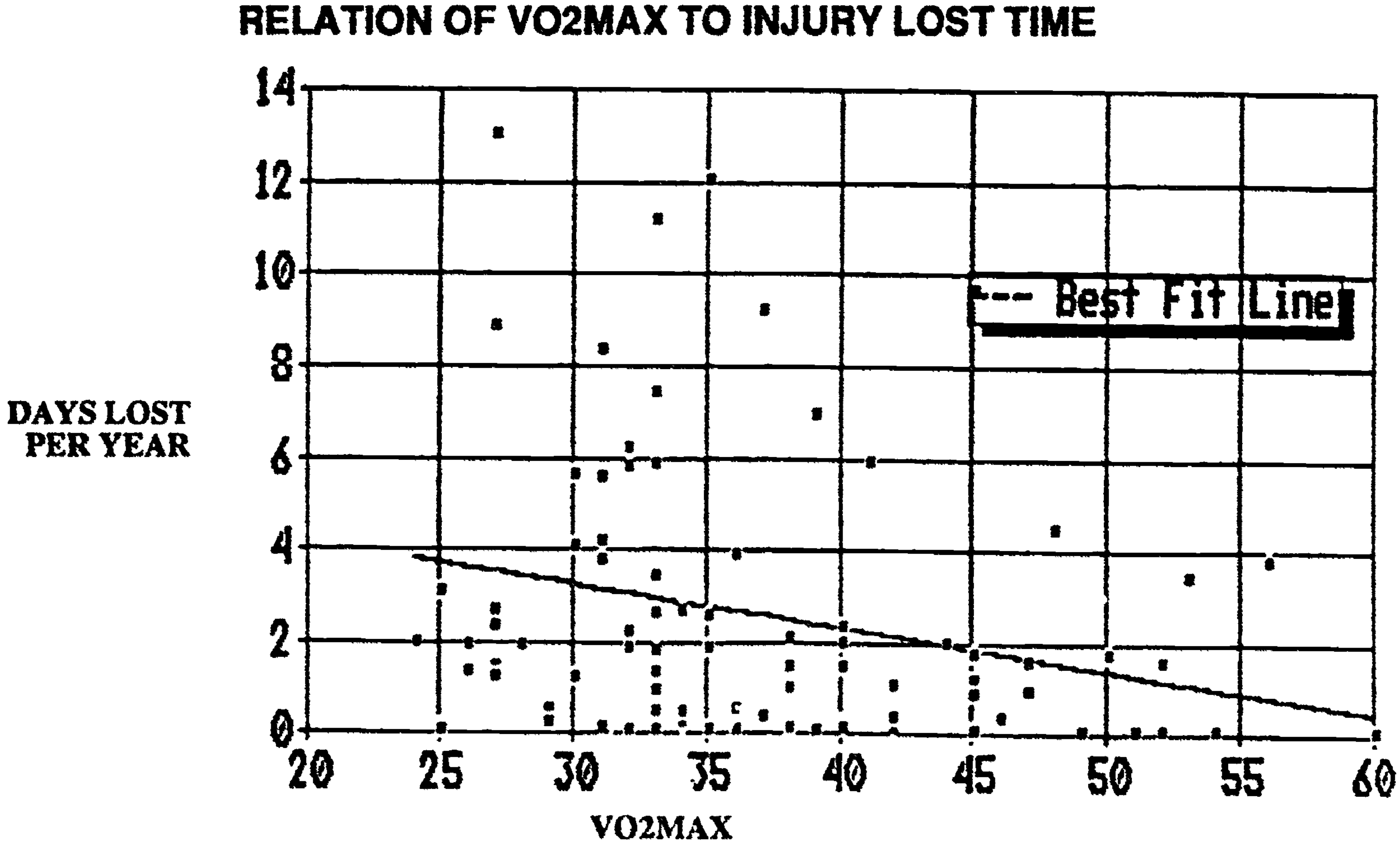


Figure 1. This figure graphically shows the relationship between aerobic capacity and injury lost time (r=-.20, significant at .05 level).

RELATION OF LIFTING STRENGTH TO INJURY LOST TIME

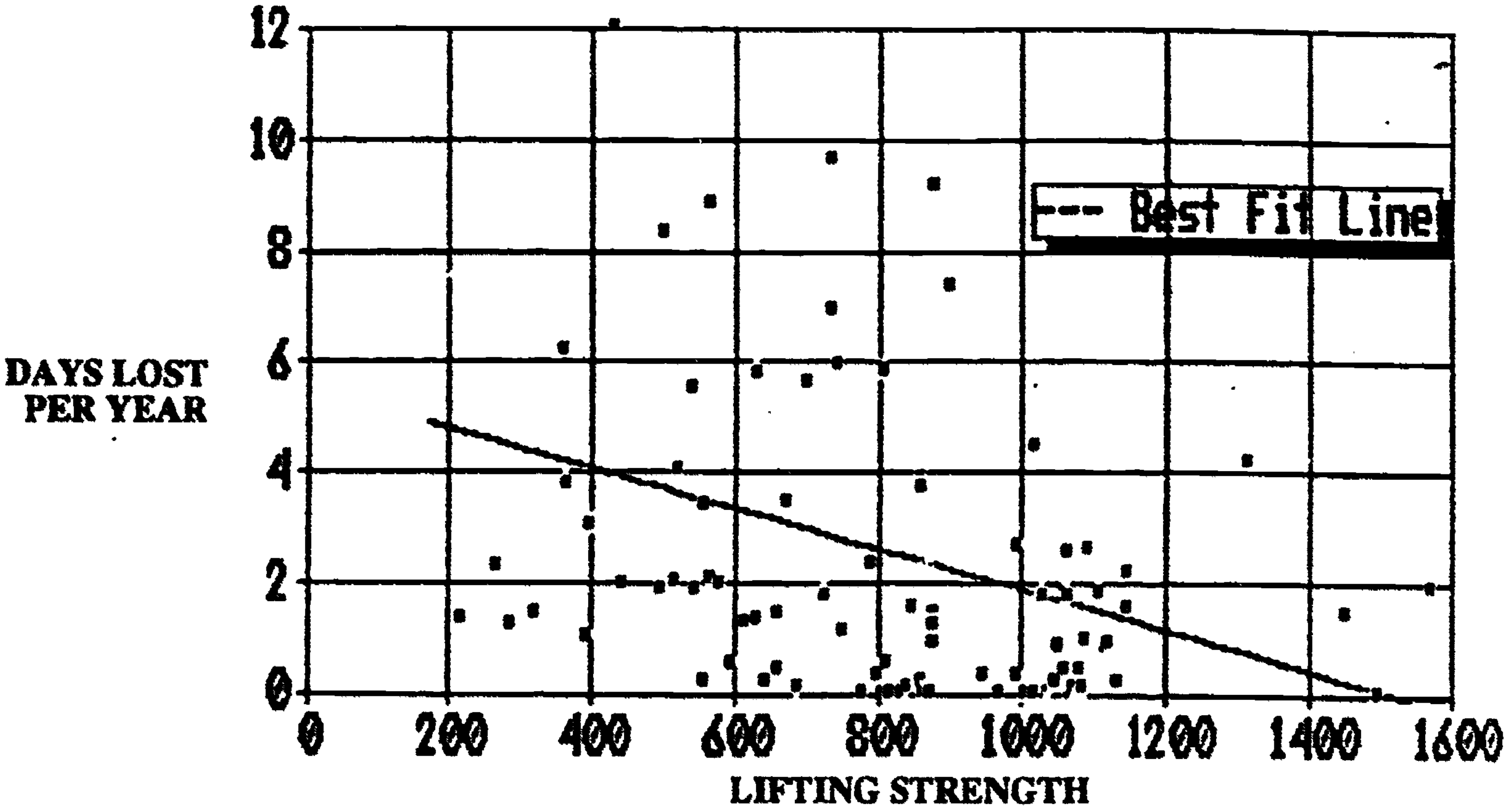


Figure 2. This figure graphically shows the relationship between lifting strength and injury lost time (r=-.30, significant at .01 level).

Male and Female Capacities to Meet Physical Job Demands

Lowering of standards, or hiring without physically testing applicants, increases risk of on-the-job injuries to men and women, but particularly to women. Figure 3 on the next page shows the distribution of tested physical capacities, as expressed in METS, (a measure of aerobic capacity), of large samples of male and female job applicants for physically demanding occupations. Also included in the table are MET requirements to safely perform various physically demanding jobs. These MET requirements are based on ergonomic job analysis.

This information clearly demonstrates that hiring without physical assessment of applicants results in increases in on-the-job injuries, particularly for women. It is important to note that individuals hired into jobs that are beyond their physical capacity are not predicted to "work themselves into shape", a common lay belief. The prediction is that they will be injured.

Due to population norm differences, women hired without physical assessment into physically demanding jobs have extremely high injury rates (in some cases, as high as 100%). Many of these injuries are permanently disabling.

Integration, Harassment, and Discrimination

In large studies of racial and sexual harassment and discrimination in various public organizations, we have consistently seen hiring practices identified as a cause of strife on the job. Generally, it will be the first item mentioned in group discussions of sources of human relations problems on the job. In one study, 289 critical incidents relating to racial or sexual harassment or discrimination were collected;

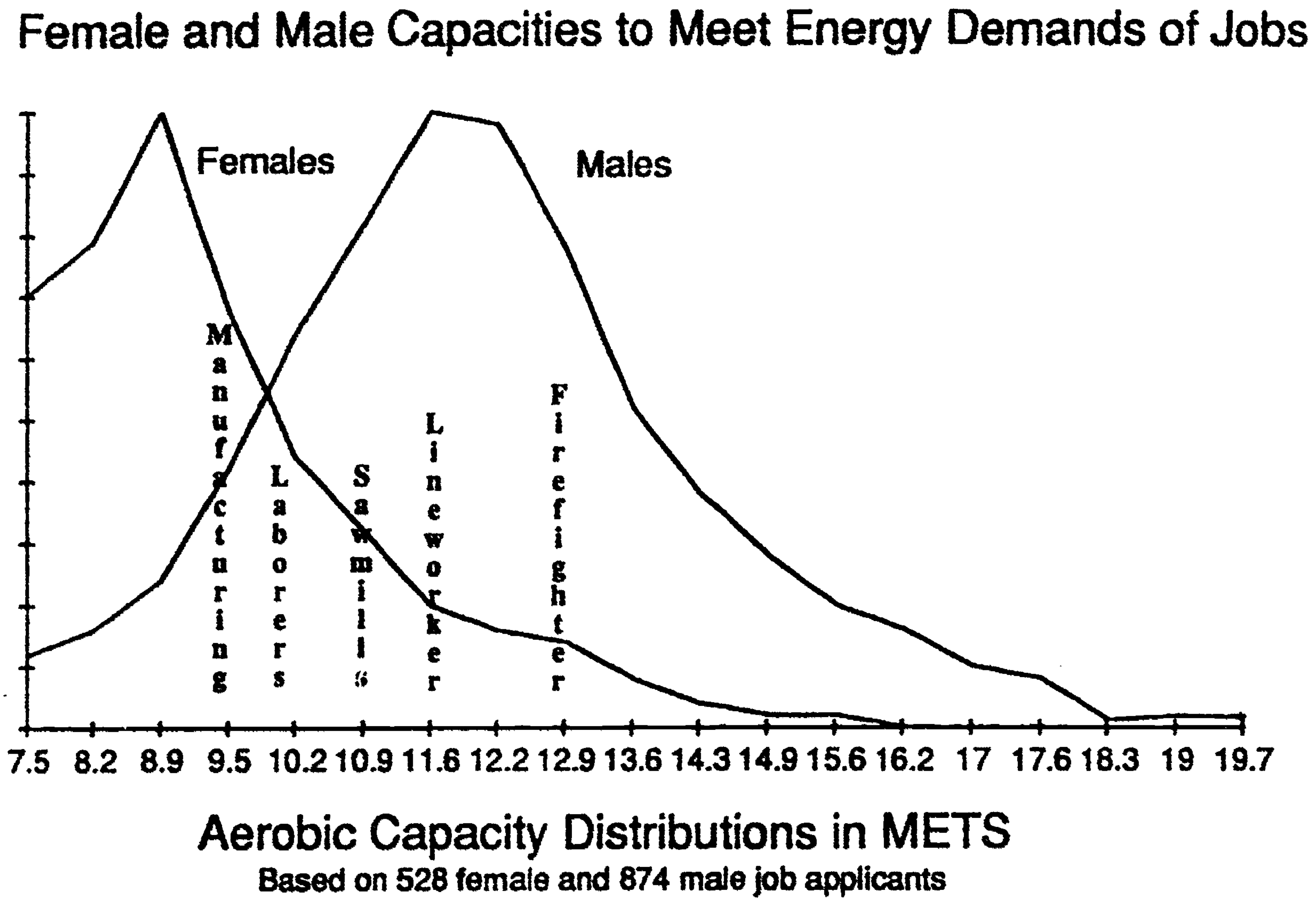


Figure 3. This table shows relative aerobic capacities of male and female job applicants for physically demanding jobs. MET demands of various occupations are showing on the graph. Manufacturing occupations generally require approximately 9 MET's; Laborers, 9-10 METS; Sawmill greenchain 10-12 METS; Lineworker 11.5 METS; Firefighter 12.9 METS. The area under the curve to the right of the occupation represents the proportion of the male and female applicant population that can be expected to safely perform the jobs.

41 incidents (14% of total incidents) dealt directly with specific criticisms of hiring and promotional practices. In another study, involving a survey of 231 employees of a public agency, there were significant differences by race and sex in attitudes about the fairness of hiring and promotional practices, with all groups (male, female, minority, non-minority) expressing suspicion of prejudice toward their group.

When protected group members are hired who do not meet physical requirements, general prejudice toward the protected groups noticeably increases, as reported by protected group members interviewed. This prejudice is difficult to reverse. Even after years of selection by appropriate standards, employees interviewed and surveyed indicate that they feel women and minorities are hired by lower standards and are therefore responsible for lower work quality, productivity, and job safety. The stigma of initial indiscriminate hiring, and consequent failures, labels qualified protected group members well into the future.

Effects of Research

In physically demanding occupations, human safety and co-worker acceptance of non-traditional workforce depend on the selection of qualified individuals. Projects designed to measure and set appropriate standards must be conducted with great sensitivity toward job incumbents. The following issues must be considered and addressed.

- Physical standards projects focus organizational attention on emotional topics. Factional rifts can deepen. Those involved in conducting the research must accept responsibility to educate and control negative rumors. Objectivity must be stressed.

- Employees fear they may not meet new standards. They fear results will not be confidential. Confidentiality of testing process and test results must be assured. This means that locations selected for physical testing of current employees must be contained, with no possibility of onlookers observing the process.

- Both males and females fear that standards will be lowered and that new employees will not be safe work partners. Women fear they will be blamed; our research indicates that this fear is well founded. Again, those involved in research must view that a major element of their work is to inform employees of the objectivity of methodology. Recommended standards must be well supported by strict, technical documentation of work demands. It should not be left to the protected group members to have to try to explain the project to others.

- The main goal for women working in non-traditional jobs is to be assimilated. Any project that singles out women on the job, even if it is for the benefit of women in general, creates difficulty for the women who are currently on the job. Researchers must be sensitive to women's concerns, preferably meeting with them confidentially as a group. Most concerns can be addressed, if they are brought to light early in a project. The women can tell the researcher what precautions must be taken and what education must occur.

Research Concerns

Women, men, minorities and non-minorities may have compelling reasons to avoid participating. Inflation of job demands is a political reality, sometimes open, sometimes covert. The researcher should

- show respect for all factions and attempt to gain their cooperation based on safety considerations,

- attempt to get information from more than one source,

- solicit information from women on any method variations they may use to modify the work procedure or use different muscle groups.

Ergonomic Considerations

When physical demands become too great (13 METS or greater), employers are forced to make job changes because they are unable to fill positions. Most physically demanding jobs are designed around the capacities of the adult male workforce.

The implication is that employers can and do modify jobs to fit the existing workforce. Some jobs can never be fully modified because external conditions are uncontrollable. Firefighting is a good example. Most heavy jobs could be modified to admit a more diverse workforce. These modifications can be costly and may take years to accomplish, but the result will be greater ability to make use of the shrinking labor pool, greater worker safety, and increased productivity.

Recommendations to lower physical job demands and ergonomically improve working conditions should be an ongoing effort of industry. However, this goal should not be used as a reason to desist in physical testing. Physical testing addresses the current working conditions, improving working safety, and preventing injuries. Even when jobs are lightened, there will still be those who are better able to perform the job than others. At that time standards should be changed to reflect new conditions.

Summary

- Employees performing jobs beyond their physical capacities are at great risk of injury.

- Due to physical size differences, if workers are hired without benefit of physical abilities testing, women will suffer far greater number of injuries on the job than men.

- When only physically qualified individuals are hired, all groups have similar, low injury rates.

- When protected group members are hired who are not capable of performing the job, seemingly irreversible negative attitudes regarding those protected groups develop among employee groups.

- Projects to set physical standards can be controversial and can cause trouble for incumbent protected group members.

- Researchers must accept responsibility to control rumors by educating the workforce with regard to merits and methodology of physical standards projects.

- Employers should not be forced to hire people who are physic ally incapable of safely performing jobs.

- Employers should work toward making more jobs more accessible to a greater percentage of the population.

Data referenced here comes from studies conducted by Carla Swander, Oscar Spurlin, Ph.D., T.L. Doolittle, Ph. D., and Barton N. Daniel, M.S., all based in Seattle, Washington.

RECENT DEVELOPMENTS IN JOB ANALYSIS RESEARCH
ADDRESS PRESENTED AT THE IPMAAC CONFERENCE

June, 1989

EDWARD L. LEVINE
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF SOUTH FLORIDA

I. FOCI OF THIS SESSION

A. Coverage of my past experience and research on job analysis.

B. Overview of selected hot issues in the job analysis domain.

C. Discussion of a small set of eminently researchable topics in this area.

D. Participants are encouraged to raise questions and make comments.

II. INTRODUCTION

A. Interests in, and familiarity of participants with, job analysis.

B. What is job analysis? (Two definitions of the terms job and job analysis.)

C. How I came to spend a major portion of my professional career in this area.

(1) Origins of research at the State of Arizona's Personnel Department

(2) Major aspects and outcomes of that research (Emphasis on "macro" as opposed to "microscopic" research; follow up studies; book on job analysis; new methods (C-JAM+B-JAM)

III. JOB ANALYSIS IS IN ITS ASCENDANCY

A. New books and chapters (e.g., Handbook of Job Analysis; forthcoming chapter by colleagues Spector, Coovert and Brannick to appear in International Review of Applied Psychology in 1989).

B. New Job Analysis Methods.

(1) Job Element Inventory ("People's PAQ"; devised by Cornelius and Hakel)

(2) Threshold Traits Analysis (Lopez and colleagues; tasks, demands and 33 traits)

(3) C-JAM (Levine, Ash and colleagues)

(a) Job related language; meets legal requirements; simple math demands; serves more than one purpose

(b) tasks like FJA

(c) KSAO's (knowledges, skills, abilities and other personal characteristics) like Job Elements

(d) scales are drawn from literature and piloting

(e) products are rated inventory of tasks, KSAO's and actions to take in using results

C. New Research.
 (1) Research is overwhelmingly oriented toward practice, as opposed to theory
 (2) Two basic themes in practice research
  (a) R&D efforts to build new methods
  (b) macro and microscopic studies comparing methods on outcomes like interrater reliability, cost, quality of information
  (c) an example of a unique research project that went beyond methods and looked at job analysis systems (Handout 1)
 (3) We need more research that is theory oriented
 (4) An example of theory driven research (Handout 2)

IV. SELECTED HOT ISSUES AND RESEARCH QUESTIONS
A. Computerization of Job Analysis (Coovert is working on expert systems).
B. Job Analysis of Team Tasks
 (1) Question of what descriptors to use
 (2) Question of compatibility among members makes specification of KSAO's difficult.
C. Validity Generalization and Job Analysis
 (1) Continuum of thought from point-for-point correspondence to broad categorization on a variable like complexity
 (2) I am for analysis of the more comprehensive type.
D. Job Analysis and Change.
 (1) Is now-should be issue
 (2) Add questions to standard methods regarding change.
 (3) Be wary of changing nature of machines-persons interfaces (e.g., machines mentoring people).

V. OVERVIEW AND SUMMARY

HANDOUT 1

DESCRIPTION AND EVALUATION OF "FLAGSHIP" JOB ANALYSIS SYSTEMS*

I. The Problem(s)

A. How do successful job analysis systems function?

B. What methods are employed by such programs?

C. How are the job analysis systems evaluated?

II. The Method

A. Nine geographically dispersed organizations were carefully selected based on their exemplary job analysis functions (banking, insurance, retailing, etc.).

B. Interviews with key personnel formed the primary method of data collection. Records/reports were also reviewed. Visits lasted one to three days.

C. Topics covered included organization context, nature of the function (e.g., history, methods used), and outcomes.

III. Results

A. Functions are centralized often in a compensation unit.

B. Most sophisticated functions have two features -- highly structured, task based questionnaires and an elaborate software system to analyze questionnaire data.

C. Compensation/Job Evaluation, Job Description, Training and Staffing are purposes most often served.

D. Tasks and worker traits are typically descriptors of choice.

E. Staff of job analysis functions are highly educated.

F. Job analysis functions are successful because they have broad base of management support (not just top management), heavy investment of resources, and highly trained personnel (i.e., Ph.D.'s in industrial/organizational psychology).

G. Cost of programs we studied ranged from $150,000 to $4,000,000/year with a median of $280,000.

IV. Implications

A. More work is needed on standards for job analysis and ways to evaluate job analysis outcomes.

B. A fully integrated personnel system based on a comprehensive job analysis data base is not feasible now. However, a multipurpose approach, designed to serve several applications, is feasible.

C. Be prepared to spend heavily to develop a successful system.

*This study was conducted under a contract from A.T.&T. by E. L. Levine, Frank Sistrunk, Kathryn McNutt and Sidney Gael. It is currently in press in the Journal of Business and Psychology.

HANDOUT 2

WHAT ARE TASK IMPORTANCE RATINGS COMPOSED OF?*

I. Research by Juan Sanchez and I had two foci:

A. Understanding rater judgments of importance (number of cues; self insight; configurality).

B. How should we measure task importance (holistic versus various decomposed indicators).

II. Method

A. Four jobs drawn from two city governments (Police Officer - 5 raters; Community Service Officer - 22 raters; Librarians - 6 raters; Engineering Technicians - 27 raters).

B. Task inventories where each task rated on time spent, difficulty, difficulty of learning, criticality, responsibility, overall task importance.

C. Individual regression equations computed; self reported weights attached to tasks; cumulative regression equations; indicators of task importance compared on interrater reliability and aggregated importance judgments.

III. Results

A. Individual regressions with scales regressed on task importance showed moderate variance accounted for; few terms used; some use of configural cues; little self insight; expertise not related to configural cue use; neither universal nor job specific policies.

B. What aggregate task importance measures appear to reflect is first and foremost task criticality and secondly difficulty of learning. Relative time spent and other cues do not contribute.

C. Two best measures to capture aggregate task importance are: (1) composite of criticality and difficulty of learning; (2) task difficulty X criticality + time spent. Overall ratings of task importance and relative time spent not as good in terms of interrater agreement and aggregate task importance criteria.

IV. Implications

A. Use as many raters as possible.

B. Use criticality ratings + difficulty of learning to measure task importance.

C. Replicate and extend these results, e.g., find out who best raters are.

D. Unexpected findings; individual task satisfaction and task importance ratings are correlated at .40 (Police Officers) and .48 (Librarians). At aggregate level task satisfaction entered regression equation ahead of difficulty learning.

*Currently in press in the *Journal of Applied Psychology*.

PRINCIPLES UNDERLYING THE COMPOSITION OF HYPOTHETICAL SITUATIONS FOR ORAL BOARD EXAMINATIONS

In attempting to assess qualifications such as oral communication skills, tact and diplomacy, and judgment, an oral examination is often the selection procedure of choice. This is, in part, because the nature of the elicited response mirrors in some respects the response that would be provided in a job setting. However, it is also due to the relative economy and efficiency of the oral exam technique when contrasted with more elaborate (and administratively demanding) approaches, such as job simulations and "full-blown" assessment centers. In jurisdictions where the staff resources are limited and/or the number of candidates eligible for examination is relatively large, the oral examination process is often a necessary, but quite acceptable, compromise in selection technology.

The composition of oral exam questions often involves the direct participation of both examination analysts and subject matter experts (SMEs). While the examination analyst may possess expertise in item writing and considerable knowledge of the job in question, his/her job knowledge may well be insufficient to formulate valid test questions without significant input of SMEs. Whether or not SMEs participate in question compositon, I believe there are a number of principles underlying such a process to which one should adhere.

These principles were formulated based upon my experience with the composition of a number of exams for various county government departments over a period of years. They are not the result of formal research or an extensive review of techniques used by other jurisdictions. While there has been some minor evolution in these principles over the years based upon experience and feedback from SMEs, they remain essentially as originally drafted.

1. The situation presented should be realistic and job-related, that is, one that someone employed in the position might well face in the normal performance of his/her duties.

Explanation: Situations that are far-fetched probably do not fairly and accurately assess the abilities under study. Such questions are also likely to be perceived by candidates as being unfair and lead to challenges to the exam.

2. The question should be specific enough in detail such that every candidate has the same general understanding of the situation described. The wording of the question should include a clear indication of what moment in time the candidate is at when he/she begins his/her response.

Explanation: Failure to do the above will lead to candidates forming varying major perceptions and assumptions about the situation, some of which may not be readily apparent from the responses provided to the rating panel. This would render uniform application of the rating criteria difficult, if not impossible.

3. Present all information in the same chronological order as the applicant would likely perceive it in real life. Also, avoid providing any description of the situation after the applicant has been "placed" at the moment in time at which he/she is expected to respond.

Explanation: Provision of any further verbage other than, "What would you do and why?", detracts from the "real time" effect one is attempting to provide for the candidate. Provision of the information out of proper chronological order necessitates that the applicant revise his/her mental picture of the situation during the course of question presentation.

4. The question should yield a variety of possible acceptable responses or approaches to the problem presented. The question should be complex enough such that any one of several variations on a theme of a response would be acceptable.

Explanation: Questions which yield only one or two possible acceptable responses generally do not adequately measure judgment, understanding of human behavior, tact and diplomacy, etc. Additionally, questions which yield only a very few specific right answers are easy to convey to and be "researched" by succeeding candidates.

5. The question should not be structured such that provision of an acceptable or better response is contingent upon (or significantly aided by) knowledge that is not quite common to all applicants who meet the minimum requirements. With police promotional exams, for example, knowledge of procedures that would likely be acquired only via assignment in a specialized unit should not be essential for an acceptable response. For an exam geared to an entire rank, e.g., Police Sergeant, the focus should be on situations a Patrol Sergeant might face, since every candidate probably has had training and experience in the patrol function.

Explanation: An oral exam is not a knowledge test. It is designed to assess those abilities not well examined in a written format. Also, neither an oral nor written exam should assess knowledge that is quickly and easily acquired through experience on the job for which the applicants are competing.

6. Avoid ascribing behavior to the candidate that he/she would not certainly display in that situation. Present all the events as transpiring before the candidate without describing any response on his/her part.

Explanation: Otherwise, some applicants will deny that they would perform such ascribed behavior during their response, thereby changing the nature of the situation presented. Other applicants will not deny the behavior but will feel awkward in responding to the question because they view the behavior as being out of character for them. Either event would make it difficult to fairly apply the rating criteria.

7. The question should not be so lengthy that memorization of all elements presented becomes a major task or otherwise interferes with the applicant comprehending and responding to the question.

Explanation: The oral exam is not designed to assess the candidate's memory.

8. The question should include only words and terms with which all candidates who meet the minimum requirements would be expected to be familiar.

Explanation: Failure to recognize or understand one or more words could needlessly confuse and frustrate a candidate who might otherwise provide an outstanding response.

While adherence to all the above principles will not guarantee a valid and universally accepted oral exam question, it should go a long way towards achieving that goal.

Clarence Weathers
Personnel Examination Manager
Jefferson County Government
Office of Personnel Management

SUMMARY

In-Basket Exam For Middle Management Jobs

Paper Presented by Mary Lindsay and Carole S. Wilcox
State of Michigan

Recently the Michigan Department of Civil Service was faced with a dilemma, a classic good news, bad news situation. The good news was that the in-basket/oral appraisal examination used for professional middle-level management positions in business and administrative occupations was a sound instrument. It was accepted and highly regarded by employing agencies. The bad news was that at six applicants per day, it would take three years to examine the backlog of interested, eligible applicants.

The resolution of this dilemma was the development of an inbasket-type examination with two essential characteristics: first, a test which could be administered to large numbers of applicants (up to 5,000 at a time) and second, which could be objectively and rapidly scored in order to create employment lists in a short period of time. An original objective had been to create a machine-scoreable in-basket. This idea was dropped when it was decided that an effective, machine-scored instrument was not feasible.

A contract was entered into with a Michigan State University professor. This consultant, a measurement expert who has experience with in-basket examinations, conducted a thorough job analysis of middle-level management positions in three professional occupational categories (business and administrative, human services and engineering and scientific). The conclusion reached after the job analysis study was that regardless of specific job categories, about 1,000 middle-level managerial classifications had sufficiently similar knowledges, skills, and abilities that they could be tested using the same instrument. This allowed the replacement of six examinations using three different examination methods (oral appraisal, paper and pencil multiple choice tests, and education and experience ratings) with one: an in-basket which can be scored without interviewing the applicant. The instrument was pretested. Score reliability data were gathered. Subject matter experts were employed in the review and refinement of the test content and in creation of the final scoring key.

The test has been administered three times; once fcr each service group. Scoring is completed for two groups and currently underway for the third. Fifteen carefully trained professional staff persons have spent approximately 1,200 person hours (combined) in the scoring process; approximately 900 person hours more will be required to score the third group. The final step of each round of scoring is the standardizing of scores from the several raters. About 5,200 applicants have now taken the examination. The statistics from the first administration have been positive in terms of reliability, discrimination, and degree of difficulty. It is expected the results of the second and third administrations will follow suit.

Abstract

The Testing of Writing Ability: An Experimental Approach, the Results and Conclusions

Eleanor R. Doyle
N.Y.S. Department of Civil Service
Albany, N.Y. 12239

Background of Study

Objective testing of writing ability is the general rule in NYS Civil Service examinations. An exception has been those for the Public Information Specialist series, jobs for which writing is the chief task and the material written is widely disseminated and critical to the success of agency projects. Several hundred candidates compete for few appointments, at most 10-15. Questions of the cost-effectiveness of making multiple ratings of multiple exercises for so many candidates for such a small payoff, and nagging concern about the precision of ratings when used to rank candidates led to our using the 1987 examination for a study to determine if an objective test would be sufficient to identify those with superior writing skills.

The objective test consisted of a 40-question Test of Written English (TWE) and a 40-question Public Information (PI) section covering publicity and public relations techniques. Four rubrics were included in the TWE: English grammar, usage and punctuation, Editing and proofreading, Information presentation, and Paragraph organization. Two writing samples were required: a 300-word news release and a 150-word public service announcement. Candidates were required to pass both parts; therefore, the writing samples of candidates failing the multiple-choice test would not be rated.

Results and Conclusions

Just over half the candidates failed the multiple-choice portion, with the Test of Written English playing the major role. Most surprising was that the two rubrics testing the most basic English language skills -- English grammar, usage and punctuation and Editing and proofreading -- were the most difficult on the test, more difficult than the other two TWE rubrics, which we consider to be testing writing skill on a higher level. Apparently, one cannot assume that candidates meeting the minimum qualifications of considerable journalistic experience and/or training necessarily possess a good command of basic English. The TWE both weeded out those weak in basic English skills and halved the number of writing samples to be rated.

Correlation of writing sample scores with TWE scores, after correction for restriction of range, was only .22. The correlation between writing sample scores and the PI section was slightly higher. The highest correlation achieved, that between total multiple-choice score and writing sample score, was .43 after correction, significant, but far too low to consider using the multiple-choice test alone to assess writing ability. Obviously, the two tests were not measuring exactly the same domain.

The relatively low correlations did not surprise those of us who participated in the rating. The TWE concentrated on standard linguistic rules and practices, and those who passed the multiple-choice test were at least minimally competent in this area. While we did see occasional mistakes in sentence construction, spelling, punctuation, or grammar, the most telling and frequent flaws in the poorer writing samples were more likely to be lack of judgment and little sense of the requirements of the two types of media material, of what information to include and which elements to emphasize, or of how to appeal to the desired audience and keep them interested throughout. Those who failed the writing sample seemed to lack the ability to objectively read over what they had written from the point of view of the audience and recognize what was detrimental to the quality or purpose of the piece. Our conclusion is that for future Public Information Specialist examinations, we should continue to use both the Test of Written English to test basic language skill and weed out those lacking the fundamentals, and writing sample exercises to assess judgment and other higher order writing skills which are so critical to successful performance in these jobs.

The rating of the writing samples was still costly and time-consuming, and although we took great pains to assure inter-rater reliability, our reservations about the precision of ratings when used to determine ranked Civil Service lists remain. We will be using writing samples in the future, but are considering a new approach. The multiple-choice test would be zone-scored. Candidates passing that test would be called back to create a "portfolio" of various writing exercises, which we would not rate. Instead, for each zone currently under consideration, we would provide certified copies of the portfolios of all candidates in that zone to the appointing officers to aid in their decision-making and to allow them to choose from among the appointable candidates the ones whose writing skills are best suited to the duties of the particular positions to be filled. We have had some feedback that agency officials would welcome this opportunity.

Copies of the complete paper, which also includes brief reviews of two major College Entrance Examination Board studies on the testing of writing ability, are available upon request.

ASSESSING THE WRITING OF TEACHER CANDIDATES: CONNECTICUT'S METHOD OF HOLISTIC ASSESSMENT

By contrast to the notion that direct assessment of writing ability is the exclusive domain of stodgy, eccentric English teachers, holistically-scored direct assessment is becoming increasingly widespread. Direct writing assessments are currently administered by the Educational Testing Service for examinees ranging from high school students to teacher candidates. In addition, they are components of the General Education Development (GED) examination as well as the Medical College Achievement Test (MCAT) and a growing number of statewide assessments including the writing subtest of Connecticut Competency Examination for Prospective Teachers (CONNCEPT).

Like the scoring method utilized for other direct writing assessments, holistically scoring the CONNCEPT is purposely a recursive, comprehensive process necessary to enhance the reliability of the assessment (an area which the literature indicates has been of some concern, but can be adequately addressed). The process begins from the point of topic development and includes the following:

1. Topic or prompt development based on test specifications
2. Topic review by subject matter experts
3. Topic field testing
4. Scoring of field tested writing samples by subject matter experts called range finders and the selection of potential training papers used for training potential holistic scorers
5. Selection and scoring of additional writing samples from actual administrations for training of holistic scorers
6. Administration of holistic scorer training, which includes

practice scoring of pre-scored writing samples, review and discussion of required writing samples as well as calibration assessments and dismissal of uncalibrated scorers

7. Scoring and periodic recalibration

The foundation of holistic scoring is the evaluation of a piece of writing based on the effective communication of a whole message. Essential to the success of direct writing assessment is the establishment of clear and consistent standards for judging the writing ability of the candidates being examined. CONNCEPT examinees are provided the following standards:

1. State and stay on topic.
2. Address all specified parts of the writing assignment.
3. Present ideas in an organized fashion.
4. Include sufficient detail and elaboration to statements.
5. Choose effective words.
6. Employ correct grammar and usage.
7. Use correct mechanics (spelling, capitalization, punctuation, paragraph form, etc.).

The score scale utilized for holistic scoring varies, with CONNCEPT relying on a score scale of 1 - 4, with the following rubrics:

1 = Has very little or no control of the characteristics that the writing of a Connecticut teache· should demonstrate.

2 = Has some control of all of the characteristics that the writing of a Connecticut teacher should demonstrate.

3 = Has generally strong control of all of the characteristics that the writing of a Connecticut teacher should demonstrate.

4 = Has virtually complete control of all of the characteristics that the writing of a Connecticut teacher should demonstrate.

The number of readers for each writing sample varies as well, with the CONNCEPT requiring three independent readings for each paper. The process of assigning final scores for the CONNCEPT writing subtest contributes significantly to scoring consistency and low rating discrepancy:

1. If scores assigned are within one point of each other, they are not discrepant, and therefore, the first two scores are summed to produce a final score of 2 - 8.
2. If there is more than a one-point difference between any pair of the three scores assigned, the paper is read by the Chief Reader, a scorer with extensive holistic scoring experience who also functions as a range finder and as the trainer of potential holistic scorers.
3. Failing papers (papers with a total score of 4 or less), are "analytically scored;" that is, they are read by an independent reader for primary areas of writing deficiency, based on the writing standards described previously. These areas of deficiency are included on individual score reports. These readers have the additional responsibility of identifying those papers they believe to be passing papers and turning them over to the Chief Reader for final scoring.

The process noted has resulted in high inter-rater agreement, with the percent agreement within 1 score point ranging from .92 to .97. In addition, consistently increasing pass rates over the first four years of the CONNCEPT indicate that this method of assessment may have contributed to an increased number of CONNCEPT candidates with improved writing ability. Finally, reporting areas of writing deficiency has not only assisted remediation efforts, it has provided some insights into those areas of writing deficiency of teacher candidates as evaluated by incumbent teachers--the analytic scorers.

Norma Sinclair
Connecticut State Department of Education

The Bush Administration and the 101st Congress: What We Can Expect in Assessment and Selection Issues

This year marks the twenty-fifth anniversary of Title VII, the most important and far-reaching employment discrimination law ever passed. In the midst of a growing civil rights movement in the sixties, Title VII was created as recognition that minorities (and, as Congress would later decide, women) were being kept out of the American workplace. The statute, as it was passed and amended, requires employers to base all employment decisions on merit alone, not race, color, sex, religion, or national origin. By definition then, employee assessment and selection is the heart of Title VII. Antidiscrimination laws are about personnel decisions.

Since the dawn of the Bush Administration there has been a remarkable surge of legal developments which will inevitably affect an employer's assessment and selection procedures. The Supreme Court has made important pronouncements about employer decision-making and affirmative action. Various bills are pending in Congress which, if passed, will impose new obligations and create new prohibitions on employers trying to select and assess candidates. The new developments may further pollute the already murky legal waters of employee selection.

This paper will attempt to clarify some of the issues by delving into the background of the Supreme Court's treatment of cases involving selection and assessment procedures, exploring the impact of recent developments, and looking toward the future of personnel assessment.

A. The Watson Decision

Last summer, the Court finally resolved a conflict in the Circuits by holding that Title VII plaintiffs may challenge subjective employment practices under the disparate impact model. Watson v. Fort Worth Bank and Trust, 108 S.Ct.2777 (1988). After expanding the use of the impact theory, however, a majority of the Court made it tougher for an employee to establish a prima facie case of adverse impact. Some of the Justices would also ease the employer's burden of showing a business necessity to rebut the plaintiff's case. Some of the Justices asserted that validation has never been required, even for standardized or objective tests, and noted that an employer is not required to comply with the validation requirements of the EEOC's Uniform Guidelines on Employee Selection Procedure to show a business necessity.

II. Affirmative Action

Affirmative action poses complex questions for employers wishing to voluntarily increase minority and female representation (or those who are required to do so as federal contractors), but who do not wish to violate prohibitions against race-based or sex-based hiring found in Title VII and the Constitution.

In January of 1989, a new Court decided City of Richmond v. Croson, 57 U.S.L.W. 4132 (1989). The Court, through Justice

O'Connor, made a showing of identified past discrimination a prerequisite to upholding a racial preference. The Court held that a Richmond program which reserved 30 percent of its construction contracts for minority owned or c,erated businesses was a form of reverse discrimination in violation of the equal protection clause of the Fourteenth Amendment. Applying strict scrutiny, the court found that the city failed to demonstrate a compelling governmental interest justifying the Plan and that the Plan was not narrowly tailored to remedy the effects of prior discrimination.

III. Legislation: Trends In The 101st Congress

Many items on Congress' current agenda also have potential impact on assessment and selection issues. Pending legislation, including the Americans with Disabilities Act, the Family and Medical Leave Act, and Workforce 2000 must be analyzed in light of Watson's and Croson's lega disparate impact analysis to subjective decision making; the merging of the heretofore distinct allocation of burdens in the disparate treatment and disparate impact frameworks; the ambiguous status of necessary validation; and the uncertain scope of affirmative action.

IV. Conclusion

The movement of the courts appears to be toward more stringent proof requirements for plaintiffs seeking to show discrimination, and more stringent proof requirements for employers seeking to remedy discrimination. The legislature, on the other hand, is endeavoring to create new causes of action for employees and to require more aggressive affirmative action by employers.

Assessment and selection personnel should continue to employ validation procedures to ward off impact suits, although the formalistic approach of the EEOC's Guidelines may not be necessary. Affirmative action which does not involve preferences or quotas is an effective method of fighting underutilization without inviting reverse discrimination challenges. Employers considering selection preferences should carefully consider Croson's lessons and be sure past discrimination can be borne out with accurate, significant and relevant statistics.

Richard T. Sampson
Semmes, Bowen & Bowen

How to Screen 50,000 Applicants for 350 Jobs: A Selection System for the Start-up of a Beverage Manufacturing Plant

Background

A major corporation in the food and beverage industry was starting-up the first new plant in twelve years. From the early planning stages of the plant, there was an awareness that interest in employment at the facility would be extremely high. This interest was attributed to several causes: 1) high wage rates and excellent benefits, 2) excellent company reputation, 3) high unemployment in the state, and 4) a good work environment.

Given the anticipated high volume of applicants, a great deal of planning and discussion was given to how best handle the large number of applicants and appropriately screen them. An effective screening process became even more important given the culture that the management team was trying to instill in the new facility. A selection system was developed with the following objectives: 1) select technically competent employees, 2) select employees with the personal values which were consistent with the values developed for the organization, 3) minimize EEO\AA liability, and 4) conduct the process in such a way to minimize any negative applicant\consumer reaction.

Description of the System

Contact with the State Employment Commission was made early in the planning process for the startup. The state had expressed a strong interest to assist in the screening of the applicants. The state and county agencies had developed a model partnership program to enable them to provide efficient services of this type to local employers. This partnership enabled them to provide large resources to the project on a cost-effective basis.

The first step in the selection system was to accept and process application materials from interested parties. A special project office was established to handle all the employment activities by the state\county for the company. Company representatives worked closely with state\county employees to insure that the company's needs were met. The importance of treating each applicant as a valued consumer was instilled, and the county employees became valuable ambassadors for the company. A special database was created that included only those applicants who designated interest in this particular company.

The special project was opened fifteen (15) months prior to the scheduled startup of the facility. Applicants completed the registration materials and returned them to the special project office. Over the course of the next nine (9) months over 50,000 applicants picked up registration materials. At that point the enrollment process was terminated.

Once the applicant returned the materials, he\she was scheduled to take the General Aptitude Test Battery (GATB). The GATB has been used extensively by the State Employment Services since the 1940's and has been the subject of extensive research and validation efforts (Bemis, 1968; Pearlman, et al. 1980).

It is a battery of tests that measures a wide range of aptitudes. These aptitudes fall into three (3) main categories...cognitive, perceptive, and psychomotor. The scores in these three aptitude areas are then combined to generate a composite score that indicates the person's probable success in each of five (5) different job families.

The way the GATB scores are determined is particularly relevant. The raw scores are converted into percentiles scores showing an individual's standing within a particular national norm group. For minorities the scores reflect their standing within their particular minority group. The norming of scores in this fashion therefore, essentially eliminates any adverse impact for minorities or bias against minorities that may be present in a test.

The applicant pool consisted of 19,592 applicants who had returned materials. 14,124 went further, and completed the second step...taking the GATB test. The role of the state at this point was to refer candidates to the company for the first round of interviews. For each particular job, the applicants with the highest score in the relevant job family was referred to the company. This approach of using "top-down" referral has been recommended in the Principles for the Validation and Use of Personnel Selection Procedures (Society, 1987).

By using the GATB scores as the referral criterion, the best suited candidates are thereby referred, and the majority of the other applicants are rejected. In this situation the GATB eliminated 87% of the applicants. Since they were eliminated using a bias-free procedure, any potential liability resulting from unfair selection practices was thereby reduced by 87%.

The referred candidates were then interviewed at the special project office by representatives of the company. Each applicant was interviewed by a team of two people using a structured interview developed specifically for the target jobs. A second round of interviewing was conducted for successful applicants. All interviewers were trained in interviewing techniques, and the legal ramifications of interviewing.

Successful candidates were next scheduled for a 25 hour course at a local vocational institute. The course had a dual purpose...to provide training to applicants, but also to evaluate them further. The course was designed to provide education in the following areas: safety, quality consciousness, industrial math, communications, problem-solving, teamwork, and group dynamics. The key qualities that were being evaluated in the course were the applicant's interpersonal skills and ability to function effectively in a team environment. This was the reason for the heavy weighting of the interpersonal skills components in the course design. These components were "taught" primarily through the use of experiential group exercises. Not only is this an effective instructional mode for interpersonal skills, but also produced a great amount of observable behavior by the applicants which could be evaluated.

The course was conducted over a period of several days. Alternative schedules were designed so as to accommodate those applicants who were currently working. Supervisors from the hiring departments attended each class session to observe and evaluate the applicants.

After completion of the course. all observers then met with the department management and a representative from the Human Resources Department. The candidates were reviewed. and employment decisions made at this point. If an applicant was rejected. the Human Resources representative was involved in the decision to insure that the rejection was appropriate.

During this time the applicants were also scheduled for a pre-employment physical. This included a screening for illegal substances. Reference checks were completed at this point as well.

Discussion

The selection of employees is critical for any position. The fact that this situation was a new plant startup made this process even more critical. The above process proved extremely effective for selection of high quality, highly committed employees. It should be understood that the process was not readily accepted by all of the managers. In the planning stages. there many lively discussions regarding the merits of the various components.

The GATB was generally accepted as a legitimate component. But as time went on. the GATB screened out some individuals with whom some of the hiring managers had some direct contact. These individuals seemed to have many of the "highly committed worker" qualities that were desired. However. they did not exceed the cutoff for the GATB test. There was some concern that by using the GATB that we were "throwing the baby out with the bathwater". After further discussion, it was recognized that the GATB was not perfect for our purposes. in that. yes. maybe it was eliminating some applicants who were highly committed. But, on the other hand, it was well recognized that the GATB was identifying those applicants with the greatest aptitudes for performing the work. Even more importantly. no one could identify a better method of objectively screening out 87% of the applicants and eliminating that portion of the EEO liability.

The pre-employment training course was a second area of considerable discussion. Some of the managers felt that the course imposed too much of a burden on the applicants. The content of the course was deemed appropriate. but there was an opinion that it did not add enough to the selection process considering the potential adverse reaction from applicants. However. as time went on, and candidates were eliminated based on observation in the class. it became apparent that the class was an effective tool for doing final screening. Some of the behaviors that appeared which applicants were eliminated for included falling asleep in class. being excessively late. and not contributing in the group exercises. These were all behaviors that the managers had seen as counterproductive in their previous positions. and not congruent with the management philosophy that was developing.

In summary, the selection system developed for this plant startup operated very effectively. Although, it is too early to have objective data in the form of turnover, discipline, or operating efficiency increases. there are other positive indicators. The plant startedup on-time. and achieved full operating capacity on a very aggressive schedule. The plant has the lowest absenteeism rate in the system. It also has the lowest grievance rate...in the first ten

(10) months of operation. only 4 grievances were filed and all of were settled prior to arbitration. Visitors to the plant (including corporate officers. management from other facilities. and outside contractors) have commented on the quality of the workforce and the cooperation they have received from the employees.

Undoubtedly. there a number of factors that are attributable to the success of this particular facility. But it must be understood that one of the key building blocks in this success story is the selection of the right people. The careful selection of the people with the right values to be able to operate effectively in this environment surely had a critical impact. One simply has to ask himself...is it worth the risk <u>not</u> to go to these lengths to select the right people?...when each person is worth a minimum of $1.25 million?

References

Bemis, S.E. Occupational validity of the GATB, Journal of Applied Psychology. 1968, 52, 240-244.

Pearlman, K., Schmidt, F.L., and Hunter. J.E. Validity generalization results for tests used to predict job proficiency and training success in clerical occupation. Journal of Applied Psychology. 1980. 65, 373-406.

Society for Industrial and Organizational Psychology. Inc. Principles for the validation and use of personnel selection procedures (3rd edition). College Park. MD: Author. 1987.

Author

E. Craig McGee, Ph.D.

Flynn-McGee & Associates
324 Leeward Court
Fort Collins. CO 80525
(303)-223-9882

The Design, Development and Implementation of a Computerized Applicant Tracking System

The Recruitment & Selection section of the Palm Beach County Department of Employee Relations & Personnel needed an automated applicant tracking system. A total Human Resource Management Information System (HRIS) total system was needed in the department but in FY 1986/87, a budget had not been approved to design, develop and implement a total system. This paper is a case study on the design, development and implementation of a computerized applicant tracking system for the recruitment and selection functions in the personnel department.

Operating under annual MBO's results-oriented management was implicit and achievement of effective performance in the various Recruitment & Selection programs; production cf credible information documenting the key dimensions of program performance and results; and communication of program performance and results to policy levels in the organization was being more carefully monitored (Wholey,1983).

Further, Palm Beach County is the third fastest growing county in the United States. Each year approximately 250 new positions have been added to the personnel compliment and an average of 600-700 vacant positions have been advertised. Each year 2-3000 more applications were received and processing this large number of applications using an antiquated Lanier word processing machine became more inefficient. This was occurring simultaneous to hiring and building a team of personnel professionals in the areas of recruitment, application screening, testing, and assessment.

Additionally, the manual system prevented staff from obtaining accurate and up to date information (Kaufman, 1973). An increase in the volume of applicant data generation resulted in a decrease in prompt customer service. One of the more significant criticisms made about the Recruitment & Selection section was; 1) lengthy time involved in referring applicants. When data was collected to respond to inquiries regarding turn around time frames, status of vacancies, reasons an applicant was not referred or selected, it took hours to collect and prepare the information verbally or in report form.

Additional and more global problems also existed similar to those identified in Darany's (1984) article, "Computer Applications to Personnel (Releasing the Genie - Harnessing the Dragon)."

As a result of the aforementioned, the task of computerizing the section became the major MBO for FY 1986/87. Five categories of information were reviewed; the internal operation, external events (political issues, site visits etc.), ideas, trends on HRISs, and pressures (Mintzberg, 1973).

The manager of the section served as an implementor and project manager to initiate much of the controlled change that would occur in the organization. A written action plan that listed all the key tasks was established. The plan included; 1) target dates; 2) breaking the project into manageable chunks; 3) identifying the sequence of steps that were involved; 4) establishing the major milestones; and 5) securing participation from staff (Meyer, 1986).

Survey research was conducted. In total, nine agencies contacted throughout the United States were able to provide pertinent information. Four canned packages were also reviewed and demonstrated, but were eliminated because they were designed as total systems incorporating areas within the department the Recruitment & Selection section was not directly involved with.

It was determined that an external consultant would be hired to write the program from scratch. The earliest decision made was not to rely on in-house MIS technical staff because staff within this department was limited. Usable dBase III+ software was already installed on the personal computer within the section and could be used to develop the applicant tracking system. Additionally, the time required to develop the software in-house would take a minimum of two years which would further delay automation.

The deciding factors used to select an external consultant were; 1) the consultant was from within the state and would be available for on site visits; 2) work could be completed in the time requested; 3) training would be provided; and 4) additions/modifications would be done at a fee affordable to Palm Beach County and within the required time period (Darany,1984).

During the initial data collection process several mistakes were made; 1) "being all things to all people" occurred; 2) data that would have been nice to have were eliminated because of the additional work that would be required of one data processing clerk; 3) promises of force and time reductions did not occur immediately; 4) training required for clerical and professional staff was more than was originally expected; 5) the old system was used longer than was originally anticipated; 6) forms for data input had not been perfected and; 7) Initially two data entry clerks were not budgeted. Another error that occurred was underestimating budget requirements for the initial system and not realizing that modifications would be required so soon. As a result, implementation of the system was delayed until additional equipment and supplies were ordered in FY 87/88.

The first version of the new system called C.A.P.T.R.A.C.K.S. (computerized applicant tracking system) was actually completed in December 1987 but wasn't fully operational until February 1988. The system is now undergoing its third version.

The system is comprised of 2 Compaq Desk Pro 386 microcomputers, 2 surge protectors, 2 Nec P5 Pinwriter printers, 2 IOMEGA Bernoulli removable disc dual 20/20 meg drives with bootable option cards, 6 bernoulli 20 megabyte removable diskettes, 1 Scan-tron 1300 OMR Data Entry Terminal and 250 Auto Sheet Feeder, 2 printer connectors (parallel and serial) with 1.2 meg floppy drives, and 1 data transfer switchbox (see Figure 1 for system configuration).

The original design of the system included two separate but related menu-driven sections each operating on separate microcomputer equipment.

The "Applicant Processing" section (original and present system) contains biographical information on each applicant allows; 1) input, maintenance and display of information from employment applications; 2) input, maintenance and display of job spec codes, job titles, and pay levels; 3) creation of exam files (assembled or unassembled); 4) posting of exam results from one exam file; 5) display of exam results information from previous testings; 6) purging of old application information and; 7) backup of database and memory variable files.

The second section "Exam Processing" allows; 1) input, maintenance and display of applicant information for each exam in process or information from previous exams; 2) printing of notices to report for written testing; 3) input of exam component information and test/rating score information (Including automatic reading of test scores from a file generated by the item analysis program); 4) automatic computation of final scores and; 5) printing of exam result notices, sorry letters, interview notices, appointment letters and transcript letters.

Overall, the system accomplished the initial objectives required; 1) storage and retrieval of employment application information (eliminating the need to input the information more than once); 2) storage and retrieval of exam results information; 3)automatic generation of Test Session Notices, Exam Results Notices, Referral Lists, EEO Analysis and other notices/letters; 4) indefinite storage of the results of each examination; 5) maintenance and printing of Referral Lists; and 6) Backup of Database and Memory variable files.

The changes made to the system during the first year of operation and changes that will or have already been made, include but are not limited to the following; 1) design, development and installation of an automated requisition tracking reporting system; 2) writing/programming routines to speed up data entry; 3) revision of Mailer section to reflect changes/additions; 4) consolidation of Applicant and Exam sections to make data entry easier/faster and to allow for operation of entire system at all times by two persons.

The potential exists for greater improvements/additions. The system has enabled internal procedures to be streamlined. Feedback to the various management levels has also improved. Other early improvements were; 1) reduction in the turnaround time for referrals from a high of 7-10 days to an average of 1-3 days; 2) sharing EEO information with affirmative action staff eliminating their need to collect EEO data and; 3) reduction in time to access an applicant record for walk in or telephone inquiries.

Overall, the feedback loop process between staff ocurred on a regular basis as they learned more about how the system worked and what it could do. This resulted in immediate communication and regular feedback to management about problems or areas that could be improved. The future of the system is exciting. As the county grows so will the system. It will continue to undergo numerous revisions/updates and it will be expanded to include more sophisticated reports. Overall, the system has been a "life saver" for professional staff within the section.

Linsey Craig
Palm Beach County

ORGANIZATIONAL PERSPECTIVES ON THE SETTING OF CUTTING SCORES

SUMMARY

The issues surrounding the setting of cutting scores are both complex and technical. What is frequently not recognized is that these complex issues also include a variety of nonquantifiable organizational and individual considerations. This paper dealt with a variety of perspectives on the setting of cutting scores and offered reflections on some of the issues that arise in that endeavor.

Despite the existence of data supporting a variety of alternatives, most selection decisions involving testing are made in terms of reference to a specific, single-point critical score. Given the day-to-day imperatives of organizational life, this circumstance is perhaps inevitable; cutting scores are administratively easy to use, easily controlled, and highly objective. However, even on nationally developed and administered professional examinations, the range of cutting scores can be very great depending on the state in which one resides. Cutting scores on the same or similar tests vary across the public and private sectors, across organizations, even across jobs and units within organizations. Why such variance? Is not there a score on which we could all agree as the definition of competency on the same test no matter where or when administered? How can something as precise as a cutting score be so incredibly imprecise?

The same questions are frequently raised by other users of tests--users who often depend upon human resource professionals to perform the necessary magic to provide them the "best" or "correct" critical score. And they are dismayed when those professionals are unable to do so. The reasons for this inability lie in the fact that the process of establishing an accurate and correct cutting score is both multi-variate and indeterminant; multi-variate in that a wide array of variables must be considered simultaneously in evaluating the adequacy of alternative cutting scores and indeterminant in that the variables which must be considered interact in such a way that it is not possible to maximize all values simultaneously. Ultimately, the process requires the evaluation of a variety of tradeoffs in making decisions about cutting scores. That is, one cannot have everything and every decision has costs.

The question, "what is the correct or accurate cutting score" is, in fact, an improper and unanswerable question. The question is more properly stated as "what is the appropriate cutting score to accomplish a particular organization's specific goals under a defined set of conditions?"

For the purposes of this paper, the initial discussion focused on some of the technical and statistical issues associated with the establishment of cutting scores. Many of these issues are well known to human resource professionals. However, the issues are quite complex, and it is appropriate to begin by defining the terminology and placing the subsequent discussions in the proper technical context. Further, some of the issues

involving the establishment of cutting scores in the realm of content validated testing are of particular concern and difficulty. There are also a variety of statistical issues to be considered including evaluating the underlying assumptions on which the statistical approaches are predicated. The topics discussed include both norm- and criterion-referenced cutting scores, empirical methods and indicators, utility, and adverse impact.

Second, the discussion turned to organizational perspectives regarding setting cutting scores. These issues deal with policies and philosophies, as well as the accomplishment of organizational decision making in the establishment of cutting scores. The discussion examined the topics of organizational goals, level of ability of current employees, labor markets, affirmative action policies, and job level and type. These issues deal with the fundamental practical concerns faced daily by organizational decision makers in establishing cutting scores.

A third topic addressed employee perspectives about cutting scores, including issues having to do with questions raised by employees, the acceptance of test procedures by test takers, issues surrounding promotional testing versus pre-employment selection, and the role and viewpoint of bargaining units with respect to cutting scores.

In summary, the presentation concluded with a model for establishing and evaluating cutting scores in an organizational setting. This discussion included an examination of the complexities associated with addressing the indeterminancy of the cutting score problem in the realities of organizational life. The realities include administrative expertise, employment needs, history of selection decisions, criticality of targeted positions, organizational culture, and expectations.

S. Morton McPhail
Jeanneret & Associates

Author Index