ED 337 474	TM 017 288
TITLE	Proceedings of the 1984 IPMAAC Conference on Public Personnel Assessment (8th, Seattle, Washington, May 6-10, 1984).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	May 84
NOTE	113p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Computer Assisted Testing; *Evaluation Methods; Job Analysis; *Job Performance; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; Psychometrics; *Public Sector; Screening Tests; Test Use
IDENTIFIERS	International Personnel Management Association

ABSTRACT

The International Personnel Management Association Assessment Council (IPMAAC) is a professional section of the International Personnel Management Association for individuals involved in professional level public personnel assessment. Author-generated summaries/outlines of papers presented at the IPMAAC's 1984 conference, which focus on occupational assessment in the public interest, are provided. The presidential address is "Occupational Assessment in the Public Interest" by B. A. Showers. The keynote address is "Systems for Linking Job Tasks to Personnel Management" by E. A. Fleishman. Twenty-two papers are summarized under the following paper session titles/topics: "Further Innovations in the Use of Training and Experience Ratings"; "Non-Traditional Testing Methods and Uses"; "Law Enforcement Personnel Selection and Retention"; "Improving the Organization: Innovations in Personnel Administration"; "Psychometric Issues"; and "Cost Effective Measures". The following symposia are summarized: "Alternative Methods of Presenting Questions and Related Information to Candidates in an Oral Examination Setting"; "How To Justify Ranking When Using Content Validity"; "The Necessity for Convergence and Integration of Personnel Sub-Systems"; and "Organizational Change". Two invited speakers' papers are summarized: the Personnel Testing Council's "Performance Measures: Forms or Samples" by S. Zedeck; and the Western Region Intergovernmental Personnel Assessment Council's "Contributions of Personnel Professionals to the Bottom Line" by W. F. Cascio. One invited address is provided: "Comparable Worth" by H. Remick. (SLD)

IPMA Assessment Council

PROCEEDINGS OF THE

1984 IPMAAC CONFERENCE

ON

PUBLIC PERSONNEL ASSESSMENT

MAY 6-10, 1984

SEATTLE, WASHINGTON

Published and distributed by the International Personnel Management
Association (IPMA). Refer any questions to the Director of Assessment
Services, IPMA, 1850 K Street, N. W., Suite 870, Washington, D. C. 20006
202/833-5860.

PROCEEDINGS OF THE 1984 IPMA ASSESSMENT COUNCIL CONFERENCE ON PUBLIC PERSONNEL MANAGEMENT

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 from information available to the Publications Committee of IPMAAC. Some presenters furnished papers which generally included extensions of their remarks, while others merely furnished a topical outline of their presentations. Adequacy and detail of information available varied greatly. For a few sessions no information was available from which a summary could be prepared.

Every attempt has been made to accurately represent each presentation. The PROCEEDINGS are summaries and condensations made by the reviewer(s). Persons wishing to quote results should consult directly with the author(s). In many cases extensive bibliographies were available which had to be excluded.

PREPARED UNDER THE GENERAL DIRECTION OF:

Clyde J. Lindley
Associate Director, Center for Psychological Service
Chair, Publications Committee, IPMAAC

Credit for major assistance in the compilation of the PROCEEDINGS goes to:

Thelma Hunt, George Washington University
Ernie Long, Seattle Regional Office, U.S. Office of Personnel Management
Ronald D. Pannone, The Port Authority of New York and New Jersey

IPMA ASSESSMENT COUNCIL

T.ie INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL (IPMAAC) is a professional section of the International Personnel Management Association--United States for individuals 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
- assisting with the many legal challenges confronting assessment professionals; and
- coordinating assessment improvement efforts.

IPMAAC OBJECTIVES support the general objectives of the International Personnel Management Association--United States. IPMAAC 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
Doris M. Maye, President
Bruce W. Davey, President-Elect
Barbara A. Showers, Past President

Published and distributed by the International Personnel Management Association Headquarters:

1850 K Street, N.W., Suite 870
Washington, D.C. 20006

(202) 833-5860

Refer any questions to Sandra Shoun, Director of Assessment Services.

TABLE OF CONTENTS

Page

PRESIDENTIAL ADDRESS

Occupational Assessment in the Public Interest

Barbara A. Showers, Department of Regulation and Licensing, State of Wisconsin

When tests are used to make employment or licensing decisions, they become the visible instruments of the process of allocating economic opportunity. In times of diminishing resources and opportunities, testing is increasingly a public issue. While technical quality of tests is still a primary concern of testing professionals, sensitivity to our public responsibilities should become an increasingly important dimension of our work. The public interest is an important and currently somewhat neglected concept in occupational assessment today.

The "public" is comprised of many interests. It may be helpful to take a look at what these interests are, and their expectations of tests.

First, what do employers want from tests? The primary use of tests in occupational assessment is to identify job competence. Hiring the most competent is expected to increase productivity and reduce costs. This argument has been extended by Frank Schmidt and others to identify the potential national impact on productivity of the use of objective selection procedures. Employers also expect tests to provide neutral, scientifically accurate decisions, and provide efficient, low cost assessments for large volumes of candidates.

What does the public want? When the public participates in occupational assessment, the primary expectation is that individuals will be judged on the basis of their talent, and not on the basis of family or political connections. Tests are expected to identify personal capabilities when other indicators, such as education or social status, do not. These expectations are consistent with the democratic values of American society. This was rather dramatically expressed for example, in a 1940 U.S. Civil Service Commission report which stated, "There is no more democratic institution in this country than the open competitive examination. Under it rich and poor, society leaders and students, intellectuals and 'low brows' compete for government employment on the sole basis of character and ability to do the work. American citizens may differ in wealth, in race, or in social station, but they are equal before the law, and they receive equal treatment in the examinations of the United States Civil Service Commission."

The public expects tests to facilitate access to jobs, identify the qualified, and to provide objective and accurate assessment of skills. When tests are used in licensing decisions, the public also expects to be protected from incompetent practitioners. Those with competence to practice are expected to pass, while those who might harm the public are expected to fail.

Another publicly held concept which influences expectations of tests is the consumer concept of product accountability. The consuming public in America has come to expect that products which are marketed for consumption will not harm them, and will possess the qualities claimed for them. Along with these notions goes the concept of liability. The producer of the product can be sued for damages if the product doesn't function in the way claimed.

In a similar vein, when tests are used in licensing, the concept of restraint of trade, and anti-trust issues come into play. Licensing tests certainly restrain trade by limiting access to the profession. Government regulatory agencies are usually granted immunity from anti-trust challenge since the restraint of trade which results from regulation is in the protection of the public interest. However, a recent case has challenged the immunity of this process. More on this later.

The area of greatest conflict of public expectation regarding occupational assessment is in the area of equal employment opportunity and balancing of the work force. Tests which were expected to facilitate access to jobs have come to be viewed as barriers to jobs for minority groups. Public expectations as expressed through federal equal opportunity law and the Uniform Guidelines are that tests should produce comparable selection rates for all groups, or else the user must provide rigorous evidence of validity, and continue to search for alternatives with less adverse impact.

So far, I have attempted to identify the expectations of employers and the consuming public regarding occupational assessment, to briefly recap: employers expect tests to increase productivity, to be neutral and scientifically accurate, and to lower the costs of assessment. The public wants democratic access to jobs based on talent, protection from incompetent practitioners, product accountability, and a balanced work force.

It becomes increasingly clear as these expectations are enumerated how difficult and perhaps impossible the task of fulfilling all these public expectations might be. However, there are external forces which insist upon accountability to the public interest, and internal forces which may help us achieve it. What are these forces, and what has been their role in defining and promoting the public interest in testing?

The external forces to which I am referring are consumer advocates, the courts, and regulation through licensing.

An advocate is by definition a person who argues for a cause. The argument is by definition one sided. The "cause" in the case of testing is the plight of those who have been decided against as a result of test use in decision making. Advocacy groups have been criticizing testing for many years. The premiere consumer advocate, Ralph Nader, four years ago addressed the issues of power and pervasive use of tests in decision making and again brought the issues to national attention. The heightened public debate fueled by this report and others resulted in federal and state legislative initiatives regarding test disclosure, and in the report of the committee on ability testing of the National Academy of Sciences.

Although the Nader report was specifically focused at ETS, the issues that were raised were broader. While we as testing professionals can think of technical requirements and counter-arguments to justify many of the actions which Nader's report criticized, it is important to lay aside defensive arguments for the moment and look at how some of the issues achieved such public credibility in the context of public expectations of tests.

Two key issues are validity and security. When discussed in public debate, validity is sometimes referred to as "fairness", because the debate centers around whether or not the test is an accurate and complete measure of a person's skills. Recall that the employer expects the tests to identify the best qualified for the job, and to be scientifically accurate. So does the public. This expectation may cause the employer, or merit system, or licensing board, to rely on precise differences in numeric scores to separate the qualified from the unqualified--those that absolutely can do the job from those that absolutely can't. As testing professionals, we all know about standard errors of measurement and estimate, and that the test can only measure job knowledges and not other skills such as determination or creativity which might affect job performance. But merit system laws and licensing rules follow public expectation, and don't often allow the flexibility. So combine test scores with other measures to determine overall competence.

The National Academy of Science report recommends that test data be used with other indicators and not be used alone, due to the known imperfections of testing. However, current laws and rules envision the test as an absolute standard, and admittedly, when dealing with volumes of people, it would be chaos any other way.

However, this very real difference between the scientific image of tests and the scientific reality is a legitimate cause for public disillusionment with tests. To the extent that we cannot improve upon the scientific reality, we ought to redouble our efforts in educating the users of tests, lest we be put in the position of appearing to support unrealistic claims about our product.

The other key issue which received considerable attention from consumer advocates is the test disclosure issue. Consistent with consumer expectations of accountability discussed before, advocates borrowed from the truth in advertising issues of past years to create "truth in testing". The candidate as consumer should be able to study the test, challenge flawed items, and double check the score for accuracy. Again, American values encourage open records and fair competition based on known rules. Americans have always been suspicious of secrecy by decision makers, and in many cases, suspicions were justified. So it is not surprising, in the context of consumer expectations, that test security is a difficult concept to support.

The maintaining of strict test security has also, in my opinion, encouraged an unfortunate public image of tests as mysterious devices. If increased access to tests helps to disspell this myth, there will be some positive gain.

The chall.~ge in the test security issue seems to be to disclose as much as possible without damaging validity. As testing professionals, we know that at minimum this will require large and costly banks of items, and test content domains which are broad enough to allow such banks. But perhaps we have been too conservative about test security in the past, and not sensitive enough to public concerns about test content. At least one national test provider, the National Association of State Board of Accountancy, releases their licensing examinations after every test. Candidates simply take the test booklets with them as they leave the room. In addition, the same organization sponsors critique sessions where candidates can come to study specifics of their performance on the examination. The test developers never seem to be at a loss for new questions, and their pass rate seems to remain low and stable over time, at about 40 percent in Wisconsin. This organization is of course the exception, but I mentioned it here to challenge the assumption that some of us seem to have grown up with, that test disclosure is never feasible.

The arguments of consumer advocates have, by definition, been one-sided, and may not all have been accurate, but they have certainly had an impact on the testing community which points to the need to increase our sensitivity and accountability to public concerns.

A second external force which has had an impact on the testing community is the courts. There probably hasn't been a single year since the founding of IPMAAC that there haven't been one or more major conference presentations pertaining to legal issues. The legal activities pertaining to Title VII and the Uniform Guidelines have even played a role in establishing the interest and the need for the Assessment Council itself.

The primary legal issue which has brought us before the courts is equal employment opportunity. None of the other public interests which have been discussed up to now seem to have had the legal impact on testing that EEO has had. Perhaps this is because this is the area of greatest conflict of public expectation regarding tests. Tests which were expected to facilitate access to jobs became barriers to jobs for minority groups. The democratic expectation which had been stated in the 1940 U.S. Civil Service Report was not successful in creating a balanced work force.

Certainly the courts, and the Uniform Guidelines, have caused us to focus intense attention on test validation, differential prediction, and adverse impact, and have undoubtedly resulted in technical improvements in our testing practices. But this is one instance where,in spite of our best efforts to be responsive, we have not been able to achieve the public expectation of a balanced work force through the testing process.

It is at this point in most presentations on EEO and testing that the speaker explains that tests can't be expected to do everything. I have always been uncomfortable with this since it sounds like it could be a convenient rationalization to stop working on an unsolved problem, though I would certainly acknowledge that much work has been done in this area.

There have been a few individuals, such as Jim Outtz from Howard University, who spoke to us last year, who are still making efforts to reduce the adverse impact of tests, and, apparently with some success. Since this is a major issue of public policy, it is important that we not abandon it in frustration. Perhaps an additional constructive approach lies in the direction of innovative policies for test use.

I can't leave the topic of the impact of the courts without apprising you of a different sort of case which is currently being considered by the U.S. Supreme Court involving the potential personal liability of the examination committee of the Arizona Supreme Court. In this case, entitled Ronwin v. Hoover, the state supreme court's Committee of Bar Examiners is the licensing agency which is being challenged by a candidate who failed the bar exam.

The state supreme court was charged by the legislature with the authority to regulate the state's legal profession. The court delegated the administration of the exam to a committee of bar examiners which was composed of practicing lawyers. The committee wrote the exam and established a grading procedure. After the exam was graded the committee picked a raw score to equal 70 and scaled the scores. The committee picked the passing score to limit the number of new licensees, rather than to represent a predetermined standard of competence.

In addition, since the committee-created grading process was not specifically adopted by the licensing authority, or specifically authorized by statute or rule, the lower court found that the examining committee was not immune from liability for anti-trust damages.

If the U.S. Supreme Court concludes that the examiners could be personally liable, then a major impact of this case could be to discourage licensed practitioners from participating in examining board activities. The role of developing and approving the examination process is often willingly delegated by the board to the testing consultant because the details are difficult to understand and relatively uninteresting compared to other matters such as discipline cases. However, if the board is not immune from anti-trust allegations, then neither is the test developer. Even if the board is found to be immune, the Ronwin case indicates to me, at least, that as testing consultants, we must redouble our efforts to assure that boards understand and actively approve the processes we develop.

The second testing related impact of this case concerns the use of after-the-fact passing point setting methods. This issue may be relatively minor compared to the liability issue, and it may hinge on the particular set of circimstances in the case; but it is worth watching, too, as a possible precedent for preset passing standards.

My current experience in licensure testing has also attuned me to the concept that if testing in general comes to be viewed as a major abuse of the public interest, increased government regulation through licensing may be sought. Although regulation may be viewed by some as a way of increasing the quality of testing and the testing profession, the functions of regulation are also to limit -- to limit scope of practice, specify entry-level qualifications, screen competency through (you guessed it) a test, and provide another forum for disciplinary complaints to be heard. Our past experiences with the Uniform Guidelines and proposed test disclosure legislation give a flavor of the pros and cons of such regulation.

Interestingly, promotion of the concept of licensing most often comes from within the profession itself, by those who view the license more as a credential than limitation. I would not promote the concept of licensing for all occupational assessment professionals. There are many questions concerning how licensing might impact the profession and few clear indications of benefits to the public or the prcfession. For example, we would have to define how most of us, with our variety of credentials and work settings, differ from industrial psychologists whose work by definition requires a Ph.D. or its equivalent in psychology. Some might say we should all become licensed psychologists. While I think the Ph.D. psychology background is a useful one, I am not convinced it represents a minimum qualification to carry out occupational assessment. There are other avenues to testing skills.

The Council of State Governments has published a pamphlet entitled "Questions a Legislator Should Ask" of groups seeking regulation. Those who would consider licensure as a means of protecting the public interest should evaluate the answers to these questions for their profession. Questions include: Has the public been harmed because the occupational group has not been regulated? Are the users of services members of the general public who lack knowledge necessary to evaluate qualifications of those offering services, or are they institutions or qualified professionals who have the knowledge to evaluate qualifications? Has the occupational group established a code of ethics? Could the use of applicable laws or existing standards solve problems? Will regulation be harmful to the public? For example, will competition be restricted by the occupational group, such as prohibiting price advertising? Will the occupational group control the supply of practitioners?

While some express doubts about self-regulation, there appear to be positive signs of its effectiveness, including, for example, the Joint Technical Standards, and the ethical standards of professional associations.

We have talked about public expectations of testing, and the external forces which can be imposed to insist upon accountability to the public interest: consumer advocates, courts, and potential regulatory laws. Now let us turn to the internal forces which may help us to achieve it -- professional research, professional standards and professional associations.

Professional research is the primary creative method we have to help us achieve public expectations of tests. Research uncovers strategies for increasing the validity and reliability of tests and brings them closer in quality to the expectations of employers and consumers. It helps us to increase their precision, their job-relatedness, and their ability to increase productivity. A more valid test becomes a more fair test by more accurately identifying competencies.

But there are current inadequacies in the field of test research which make it difficult for us to develop tests which meet public expectations.

For example, we don't have clear consensus in the most important area of test research, and that is what constitutes sufficient evidence for establishing validity. The controversy surrounding the Joint Technical Standards proposed requirement of multiple forms of validity evidence illustrates this lack of consensus. How much consensus is there among us as to what test interpretations require construct validation, or how much construct validation constitutes a reasonable certainty of construct validity? Although I don't expect to be presented with a simple decision rule regarding sufficiency of validity evidence, the amount of vagueness that still exists among testing professionals on the topic is disconcerting. If we don't have a clear idea of how much evidence constitutes a reasonable certainty of validity, we can more eas ly defer to what we consider to be economically and practically feasible under the circumstances. Practical feasibility will always be a consideration, but without clear standards of proof of validity, we may begin to think that what is feasible is sufficient. It may or may not be.

A second area of inadequate research consensus is passing point setting. Competency-based methodologies have given us progress in the field, but different methods give maddeningly inconsistent results. We can currently offer some improvement over arbitrary passing point setting, but we are far from precise in this important responsibility.

Finally, the area of test bias and adverse impact has not yet yielded to our best research efforts to identify, reduce, or eliminate these effects. Because we have not yielded concrete conclusions in this area, each researcher is left to his or her own opinions as to the nature of the problem and what to do about it.

The role of the resarcher is to continue to seek closure on these complex problems which place limits on our ability to meet public expectations concerning tests. It is the most diffucult role, but one of the most important. One of its virtues, compared to the other forces I have mentioned so far, is that it is pro-active, rather than re-active. It is potentially our greatest source of strength.

Another internal force which impacts and is affected by the public interest is professional standards. The force of public interest is acknowledged in many ways in the introduction to the new draft of the Joint Technical Standards. One example states: "Recent controversies over testing make the development of these Standards difficult. The Standards do not attempt to provide psychometric answers to policy questions. However, complete separation of scientific and social concerns is not possible."

The role of the Standards, as stated by the authors, is to provide a technical guide and basis for evaluating testing practice. The underlying philosophy, when it comes to social impacts of testing, is "to advocate that . . . the necessary technical information be made available so that those involved in policy debate may be fully informed."

While the Standards do not advocate test disclosure in the same way that consumer advocates have suggested, the emphasis on availability of technical information is consistent with the public interest value of open records which allow independent verification of the quality of the test.

In addition, the Standards show sensitivity to the public interest by including requirements which protect the test-taker. For example, the chapter on test administration includes a standard which prohibits release of a person's test scores to others without the person's consent.

While most of our attention as testing professionals has tended to focus on what the Standards will require us to do, and whether or not the validity chapter represents a consensus of professional opinion, the ultimate function of the Standards will be to provide a public document of basic expectations for testing practice which can be used both to inform the public and regulate the profession.

Finally, I have referred to professional associations as one of the internal forces which can be used to help us achieve public accountability. Professional associations in general are often viewed more as self-interested than public-interested. And, in fact, except for the language in the bylaws which defines us as a tax-exempt organization exclusively for charitable, scientific, and educational purposes, our goals are all focused on advancing our professional interest, such as sharing ideas, and advising others of our position. These are not wrong objectives, but what I am suggesting here today is that we balance our goals with a sensitivity to public expectations of testing. Perhaps we need to include in our organizational goals some overt efforts to balance public and professional concerns by including the input of non-testing professionals, for example consumer advocates, in our programs and publications, and possibly even on our Board.

We should encourage diversity in our membership as well, encouraging both private and public testing professionals to join, and the broad base of occupational assessment fields.

The theme of this presentation has been that sensitivity to our public responsibilities is as important as improving test quality through research. I have attempted to identify the public expectations of tests, and some of the internal and external forces at work to help us or force us to realize these expectations. I have placed primary creative responsibility for meeting many of these expectations on the shoulders of test researchers, and pointed out some current limitations in the testing field which are preventing us from meeting public expectations. I have also suggested that there is a role for associations such as ours in balancing the public and professional concerns.

I believe that when the public views tests negatively, those views have been caused in part by our insensitivity to the impacts of tests on the public, and perhaps by our underestimating the high expectations that the public has concerning tests. We should give more thought to being responsive to public concerns. Our initiatives could include, for example, more active public education programs on test use and test taking which could be sponsored by IPMAAC, or at the individual level, more openness in providing test content information to candidates. There are undoubtedly many more ideas which could be proposed, if we put our minds to it.

It is this concept of closer attention to public concerns which I see as an important and currently somewhat neglected focus in occupational assessment today. Occupational assessment is in the public interest, and we should strive to keep it that way.

KEYNOTE ADDRESS

Systems for Linking Job Tasks to Personnel Requirements

Chair: Cassandra K. Scherer, Conference Program Chair

Address by: Edwin A. Fleishman, President, Advanced Research Resources
Organization, Washington, D.C.

The paper explores ways of describing human tasks which might improve predictions about how people will perform on such tasks. It is presented as part of a more general program concerned with taxonomic issues in the behavioral sciences. The assumption is that in the world of human tasks common task dimensions can be identified which will allow improved prediction of human performance on these tasks. The paper summarizes some of the efforts to conceptualize human tasks. It deals with some objectives, basic and applied, of developing a general taxonomy, or classification, of tasks. Then it describes some alternative task classifications examined, with examples of how one approach to classifying tasks has led to a number of applications in work situations, including applications in areas of public personnel assessment.

Much of our research in the behavioral sciences is concerned with the study of factors affecting human task performance. Such factors include different learning conditions or training methods, different physical and social environments, different motivational and attitudinal factors, or individual differences in abilities. The one set of variables common to all these areas, are those associated with the kinds of tasks that people perform. What has been lacking is a system for classifying such tasks.

Behavioral scientists in basic as well as applied fields have recognized these problems. Melton and Briggs pointed out the need for taxonomies of skills to deal with the expanding universe of knowledge in engineering psychology. Paul Fitts called for a taxonomy which identified important correlates of learning rates, performance levels and individual differences, equally applicable to laboratory tasks and to tasks encountered in industry and in military service. Robert B. Miller called for a behavioral taxonomy related to the generalization of characteristics of task performance, which would enable the task analyst and training designer to find a common ground in the research literature. In spite of these earlier expressions of concern, until recently, few systematic attempts at taxonomic development have been undertaken.

Potential Uses of Task Taxonomies

A number of ostensibly disparate problems can be viewed in a new light by applications of such a taxonomy. Starting with basic research implications, some are listed.

1. Conducting literature reviews. Our first encounter with classification takes place when we try to locate literature relevant to our research. We are faced with the problem of locating and matching descriptors of human task performance in the literature with our

own particular terminology. Are we dealing with the same or a different class of human performances?

2. Establishing better bases for conducting and reporting research studies to facilitate their comparison. After completing our research, we will again confront the same problems in relating the results of our experiments back to a body of experimental or theoretical knowledge; this leads to a second application. A comprehensive classificatory system should aid in disclosing the reasons why studies can or cannot be compared. A taxonomic system could at least provide some guidelines for improving the conduct and reporting of research.

3. Standardizing of laboratory methods for studying human performance. A critical problem in the experimental study of factors affecting human performance is the lack of standard tasks and measures. One spinoff from research on taxonomic questions can be the specification of standardized tasks which are diagnostic and reliable measures of defined human functions.

4. Generalizing research to new tasks. A human performance taxonomy should assist in extrapolating from previously attained research results to new tasks. For example, the effect on performance of a given environmental factor, such as high temperature, on Task A may be known, but will this hold for Task B or Task C? A useful taxonomy would tell us if these tasks are in the same or different categories as a basis for generalizing from Task A to the other tasks in the same category.

5. Assisting in theory development. There are many points at which taxonomic development supports theory development. The success of a theory primarily depends upon how satisfactorily the theory can organize the observational data of the science. In developing theories about human performance, we need concepts which will help us classify these performances.

6. Exposing gaps in knowledge. A taxonomy can help expose gaps in the body of knowledge regarding human performance. By delineating categories and sub-categories of human performance, a taxonomy makes much more evident where extensive research has been done, and conversely, where it has not been done.

In addition to these general basic uses, the ways in which a taxonomy would be useful in applied and practical areas of human performance include the following:

1. Job analysis. A taxonomic system utilizing appropriate general descriptors can help establish the similarity of new and different jobs and can group jobs into families having similar personnel requirements.

2. Person-machine system design. The planning and allocation of functions to man and machine requires the making of decisions about human performance. An important input to such decisions should be the category of performance with which one is dealing, and the categories of the various factors which can affect that performance.

3. Personnel selection. In order to effect the most suitable match of people to jobs, data about the task dimensions of the job and about the characteristics of personnel must be available. A useful taxonomic system would include concepts linking the characteristics of job tasks, their performance requirements, and the capacities measured by selection tests. We'll say more about this later.

4. Training. Application of the "principles of learning" to training would appear as desirable, but is quite difficult in practice because there is insufficient information about the categories of human task performance within which different training methods are effective. The problem is one of developing a classification system which will match those training techniques found effective with particular categories of skill.

5. Performance measurement. Many investigators have recognized the need for "standards" of human performance which can serve as points of reference for the effects of experimental variables and program interventions. The development of a taxonomy of human task performance would provide the foundation for new valuable measurement techniques.

6. Development of retrieval systems and data bases. An entire field of information science has been developed, with associated computer systems for the storage and rapid search and retrieval of scientific information. The efficiency and utility of such systems could be enhanced if the information about factors affecting human performance were indexed according to the class of human performance involved.

Having stated these potential uses, we can see the diverse implications of advances in this area. They provide a set of objectives to guide future taxonomic development and a set of criteria against which the utility of future taxonomic development can be assessed.

Some Issues ir Classifying Human Performance.

In working toward these objectives, the author examined the experience of other sciences, where taxonomic development has a longer history. This, it was felt, would help us understand the relevance of these issues to taxonomic development in the behavioral sciences. The review underscored the need to establish the purpose and method for developing a classification system before one attempts to classify.

The author found a diversity of definitions of tasks ranging from the totality of the situation imposed on the subject to specific performances required. These different definitions led to different models and rationales for describing and classifying tasks. Most definitions treated tasks as consisting of interrelated processes and activities. Our conclusion was to adopt definitions that permitted the derivation of terms that reliably describe tasks and distinguish among them. These derived terms can then provide the conceptual basis for classification. From this analysis, four primary bases for task classification were derived. These bases, were:

behavior description approaches (e.g., handling objects, analyzing data), behavior requirements approaches (e.g., problem solving, scanning), ability requirements approaches (e.g., spatial-visualization, verbal abilities) and task characteristics approaches (e.g., type of display, instructions, goals).

The behavior description schemes classify human tasks in terms of overt behaviors, based on observations of what individuals actually do while performing a task.

The second approach, called the behavior requirements approach, emphasizes the task behaviors in terms of the types of inferred processes required to achieve certain criterion levels of performance. The employee or human operator is assumed to possess a repertoire of processes (or functions) that intervene between stimulus events and responses and these can be codified.

A third conceptual basis, which we have called ability requirements approach is, in many ways, similar to the behavioral requirements concept. Abilities are relatively enduring attributes of individuals. The assumption is that specific tasks will require certain abilities and those tasks requiring similar abilities can be placed in the same category. Abilities differ from behavior requirements (or functions) primarily in terms of concept derivation and level of description. A primary source of information are experimental factor-analytic studies of individual differences in task performance.

A fourth approach, called the task characteristics approach, is predicated on a definition of a task as a set of conditions which elicits performance. These conditions are imposed on the individual and have an objective existence quite apart from the activities they may trigger, the process they may call into play, or the abilities they may require. Appropriate descriptive terms are those which focus on the relevant task stimuli, instructions, procedures, response characteristics, and goals.

From reviews of the earlier work, it was concluded that neither highly specific nor highly general categories are likely to be the most useful in generalizing principles across tasks. Also, it was found that little empirical evaluations had been made of the extent to which these various descriptive systems could improve prediction and generalization about factors affecting human performance. The arguments for and against various approaches, and the preliminary conceptual development gradually convinced us that more than one provisional approach was needed.

The decision was to develop a number of alternative taxonomic systems, based on different rationales about common factors in task performance. This may, in retrospect, appear obvious, but at the time it was an insight which provided an advance towards the solution of some taxonomic problems. Some of these approaches were essentially empirically inductive, while others involve testing of a priori theoretical formulations.

Evaluation of Taxonomic Systems

We gave major attention to the development of criteria and evaluative systems for testing the reliability, validity and utility of these approaches. Another consideration included defining the requirements for data bases to be used in evaluating the capabilities of the various taxonomic systems to integrate the experimental literature. We felt that the development and validation of any taxonomy of human performance is highly dependent on the data in the existing literature. Consequently, attention was given to an information system to provide access to the research relevant to the classification of human performance.

Ability Requirements Approach

The approach which has received the most development and most extensive evaluation are extensions of the ability requirements approach, in which tasks are described in terms of the human capacities required to perform them effectively. Tasks are categorized according to the common abilities required. The abilities on which the system is based were derived from empirical studies on the interrelationships among performance on a wide variety of tasks, including the sensory, cognitive, perceptual, motor, and physical performance areas. Individuals performing in factor analytic studies or other clustering methods form the initial bases for these dimensions.

In reviewing our studies, it became apparent that in defining these ability factors we were really linking up a great deal of information about task characteristics and ability requirements. It was possible to state a number of principles relating task characteristics to ability requirements. For example, it was possible to say that an ability called "Multilimb Coordination Ability" was common to tasks involving two hands, hands and feet, etc., in operating equipment, but did not extend to tasks in which the body was in motion, as in athletic skills. We could show that there was an ability common to simple auditory and visual reaction time tasks but requiring choices between responses or stimuli shifted measurement to another ability called "Response Orientation." It was shown that it is not too useful to talk about strength as a single physical ability; in terms of what tasks the same people can do well, it is more useful to talk in terms of at least three general strength categories which may be involved in different ways in a variety of physical tasks.

The following illustrates the use of the modified Ability Requirements Approach in the context of personal applications:

Predicting Learning and Performance Levels

We attempted to use ability concepts developed to predict various learning measures and other aspects of task performance. In general, these studies with a variety of practice tasks showed that the role of various abilities at different stages of learning could be traced. Some abilities were predictive of early learning and others predictive of later learning. Thus, some of the taxonomic criteria proposed by Fitts were met by the ability concepts, since they were shown related to learning rates, performance levels, and individual differences.

Development of Standardized Tasks

As another illustration, we have developed laboratory tasks representative of the various categories in the ability requirement taxonomy. Such "standardized tasks," representing the ability dimensions, have been used in laboratory studies of various factors affecting human performance. Thus, we have studied the effects of a variety of drugs and dosages on measures of a variety of reference ability tasks. An illustration is provided by our results with a given dosage of the drug scopolamine. We obtained different effects according to the tasks performed within a variety of cognitive, perceptual, and motor areas; that is, some abilities within each area were more affected than others by the same dosage of the drug.

We have also conducted similar studies on the effects of different noise stressors, where the principal finding was that intermittent moderate intensity noise affects performance on tasks emphasizing some abilties (e.g., reaction time) but not others (e.g., rate control).

Integrating Research Data

The ability classification approach was also evaluated in terms of the capacity to reorganize and integrate areas of the human performance literature in meaningul ways. Thus, improved generalizations about factors affecting performance in long-term monitoring tasks were obtained when the data in this literature were replotted according to the task's ability categories.

Measurement Systems

One of the striking findings in our review of the factor analytic literature was the difficulty in moving from the factor analyst's definition to a more operational definition which could be used reliably by observers in estimating the ability requirements of a new task. A large effort in our program involved the successive refinement of such definitions to improve the utility of these concepts in describing tasks and the development of measurement systems.

Recent Applications

In terms of the criterion of utility the ability system for describing human tasks has found application to a variety of applied problems. As a method of job analysis and test development it has been employed in a range of studies including determining the requirements of firefighters, grocery warehouse clerks, telephone line workers, probation officers, refinery workers, Army and Navy occupational specialties, accountants, inspectors, maintenance personnel, etc. Tests selected to map on to the abilities identified have been shown to have criterion-related validity. The abilities analysis method of job analysis is particularly relevant to issues of content and construct vaidation, since the method provides the basis for demonstrating the job relevance of the ability tests selected and their linkages to critical job tasks.

We have used these methods as a basis for setting standards for assessing job performance. Specific tasks comprising a job can be evaluated with respect to their requirements for various ability factors. Tasks rated highest on the different scales across a wide variety of occupational specialties are selected for work sample or criterion referenced tests. Individuals who can perform these tasks can be assumed to be able to perform all other job tasks rated lower on the same ability scales. For example, for the job of truck driver, changing a 50 lb. tire was the task rated highest on the Static Strength tests. Individuals who would perform this task could perform the other tasks falling on the Static Strength Scale. The scales provide a means for identifying the relevant tasks covering the abilities required by the job in a cost-effective, objective manner.

A more recent development is the use of these methods in the area of setting medical standards for physically demanding jobs. The medical examination, administered by physicians, is coming under increasing scrutiny for job relevance. Medical screening is often done without clear enough information about the job tasks and requirements. Current work underway in our own program has attempted to link the ability requirements of job tasks with the diagnostic procedures utilized by examining physicians. The abilities taxonomy allows integration of such requirements across a great many jobs. Working with specialists in occupational medicine we have developed physicians' manuals in which the disqualifying symptomologies at each level of each physical ability requirement are provided.

These levels were established at the task level across a great variety of tasks but the taxonomy allowed their integration into the several ability categories. Using the manual developed by this process the physician notes , the rating of each job on each ability and can relate the symptomologies observed in job applicants to the job related guidance provided in the manual.

The system described also appears to be useful as a method for classifying, grouping, and indexing jobs in terms of common ability requirements. Thus, diverse jobs involving many different types of tasks have been grouped according to the common abilities needed to perform them effectively. We now have developed a data base of thousands of job tasks, whose scale values on different abilities have been determined. As new jobs containing similar tasks are analyzed their ability requirements can be estimated from this data base.

Summary

I have described a particular programmatic effort, with some recent developments and applications. We are encouraged that taxonomic systems can be developed and that a taxonomy linking abilities and task characteristics meets a number of criteria across a variety of basic and applied areas of psychology and personnel research.

Recent empirical work suggests that the most useful set of primary categories in contemporary taxonomy appear to involve a rather large and steadily increasing set of categories. We needn't feel self concern about this since this was shown to be true in the field of biology as well as in the fields of human learning and performance. The increasing fractionation of categories, while perhaps complicating life, is consistent with empirical work in the interrelationships among human task performances.

More importantly, if nature is more complex than we would like it to be, we need to take steps to organize and conceptualize it in ways which make it more manageable. Tnus far the results are encouraging that a system for linking job tasks and ability requirements can solve a number of important problems in the area of personnel assessment.

PAPER SESSION

Further Innovations in the Use of Training and Experience Ratings

Chair: Sally J. Brauer, U.S. Postal Service, Washington, D.C.
Discussant: James C. Johnson, State of Tennessee, Department of Personnel

Computerized Rating of Training and Experience

Ernie Long, Seattle Regional Office, U.S. Office of Personnel Management

In the U.S. Office of Personnel Management, evaluation of training and experience is one of the primary tools for determining applicant eligibility and rank according to qualifications. As our budget has gotten tighter over the last few years, the staff available to evaluate applications has decreased. Applicant volumes remain high. We therefore needed to find more efficient methods to handle the workload.

We have two needs in the area of applicant evaluation. First, we need to determine minimum eligibility for work in the occupational series and at the grade level for which the applicant is being considered. It basically consists of determining whether an applicant has enough months or years of the right type of experience to meet the minimum eligibility criteria. Second, for all applicants who meet the minimum criteria, we need to rank order them according to relative merit.

The applicant rating process divides itself fairly naturally into these two parts. Normally, different criteria are relevant for determining minimum eligibility for a position than are relevant in ranking. The knowledges, skills, abilities, and personal characteristics (KSAPs) required to be minimally competent in a job are often not the same ones that distinguish superior workers and thus may not be appropriate to use in ranking applicants. This is a critical psychometric point and one I would like to talk more about later.

The Computerized Rating Procedure

At present, most of our T&E evaluations are done by a professional rater. It involves manual review of an SF-171 (application form). Depending on the occupation being rated, it takes approximately 20-30 minutes per application to determine minimum eligibility for all the various grade levels and approximately 30 minutes per application to "fine-rate" (assign a score of 70-100).

Our computerized rating process, at the moment, focuses only on the second type of rating, that is, assigning a score indicative of relative merit to applicants who have been found to meet minimum eligibility criteria. Our computerized rating procedure has reduced this rating time from 30 minutes of a professional's time to less than 5 minutes of a clerk's time. Depending on whether applicant data are key-entered by a clerk .. a terminal (or PC) or read from optically-scanned test answer sheets, the cost reduction over the old manual rating process ranges from 76% for clerical key-entry to 98% for optical scanning.

The basis for the rating is a supplemental application form, completed by applicants, as shown in the handout sample for our prototype for Electronic Technician. Applicants self-rate their skill level on a set of 200-300 specific job tasks using the "skill level" scale shown. The basic rationale is that the applicant who claims (and is able to substantiate) the highest level of skill on a sufficient number of tasks gets the highest rating.

There are seven specialties within the general field of Electronic Technician work (computers, communications, radar, etc.). Applicants indicate their skill level on a set of GENERAL tasks that are common to all specialties and the tasks for each specialty for which they wish to be rated. Their final rating is a combination of their score in the specialty with their score on the GENERAL tasks. This scoring rationale reflects our belief and policy that the best qualified applicant will be the one with the best preparation for the general field of Electronic Technician work and the best preparation for the particular specialty.

Responses ABCDE for each task are assigned point values of 0-4 and the person with the highest skill level per task marked gets the highest rating. Tasks are presently all given equal weight although future research may suggest some advantage to differential task weighting.

Exaggeration Checks

Since rating is based on self-assessment, the potential problem of exaggeration must be considered. We have built in several features designed to discourage or catch exaggeration. One is the "Block II" information which asks the applicant to substantiate his/her claimed level of ability by indicating the experience or education which gave him/her the level of skill claimed. Second, applicants are told that former supervisors or teachers may be contacted to verify claimed skill levels. These steps may discourage some exaggerators.

Built into the computer program are some additional exaggeration checks. Since our objective with computerized rating is to approximate as closely as possible the judgment of the human rater, we asked the rater what the signals to him were when he manually rated an application. His experience was that an applicant would normally not be able to substantiate a skill level claim of D or E for more than half the tasks in any sub-specialty, mainly because of the nature and breadth of the tasks in the inventory. So the third exaggeration check is that the computer will flag any applicant who marks too many high-level responses. The application is then given to the rater to seek information to corroborate the claimed skill level either within the application itself or by getting additional information from the applicant.

The fourth exaggeration check is also done within the computer program. We also know from experience that applicants tend not to be able to become well qualified in more than two specialty areas. So if a score is too high in more than two specialty areas, the computer flags the application for the rater to review.

User Feedback

Follow-up and feedback on the quality of applicants being certified is at this point anecdotal. The federal agencies to whom we provide rated applicants have uniformly expressed pleasure and confidence in the quality of the applicants they have received. Agencies are particularly pleased by the level of detail in the information they have about an applicant's experiences, as reflected in the applicant's responses to the task inventory which accompanies the certificate. Responses to the task inventory are routinely used as a basis for interview questions for certified applicants.

Issues for Consideration

In going to computer-assisted rating, one of the first commitments that must be made is to task-based examining vs KSA-based examining. For many people, myself included, this may be a disquieting transition. There are some fairly substantial tradeoffs between task-based and KSA-based examining.

Self-assessment Issues

If one is able to satisfy one's self as to the acceptability of a task-based examination and job analysis process, the next issues that are critical to the viability of the task-based examining process are the issues involved in the concept and process of self-assessment. Can applicants reliably rate themselves on task-based dimensions?

I have prejudiced the answer to that question by the phrasing of the question itself. If the dimension being self-rated is performance of a task as opposed to possession of an ability. I believe there is immediately a higher potential for accurate self-reporting. Performance of tasks is a more objective phenomenon than possession of abilities, even such fairly tangible abilities such as "typing speed."

To the extent that task-based examining devices can use such self-report scales, in lieu of "self-assessment" scales, I believe that, for this purpose, self-reporting will be more accurate and reliable. The handout sample of the self-report scale for Nursing Assistant may help to clarify the differences, especially when contrasted with the Primoff-inspired (Primoff, 1975) scale we used for Electronic Technician which, especially in skill level E, still has a trace of subjective "self-assessment" where the applicant is asked to decide if she/he is an expert."

The other key to maximizing the reliability of self-report data seems to be, as other writers have adequately documented (Primoff, 1980), the importance of having the clearest possible definition of the domain on which people are being asked to evaluate themselves. I believe that this points to yet another positive feature of task-based examining. The domain being self-reported is performance of a task, which seems much easier than to self-assess on a construct or KSAP (such as "filing ability") whose definition may be more subject to interpretation from one person to another and thus lead to reduced reliability in the rating.

Conclusion

There are many other aspects of task-based, computer-assisted rating that are worthy of discussion and investigation, such as interpersonal differences in self-reporting tendencies. I would be very interested in knowing your experiences or ideas in any of the areas I have mentioned.

The Activity/Achievement Indicator: A Possible Alternative to the Behavioral Consistency Method of Training and Experience Evaluation

Ronald A. Ash, University of Kansas, School of Business

Abstract

This paper describes the development of a new type of selection procedure, the Activity/Achievement Indicator (A/AI), and compares it with the behavioral consistency method of T&E evaluation.

The goal of the behavioral consistency method is to rank-order applicants on the basis of the kinds of achievement behaviors that are required for superior performance on the job in question. In general, the method begins with the generation and evaluation of job-related knowledge, skills, and abilities (KSAs) by subject matter experts (SMEs). The KSAs are combined into major achievement dimensions, usually from five to seven. For each major achievement dimension, applicants are encouraged to describe in detail at least two past achievements which best demonstrate their capabilities. Applicants are asked to include the following information for each achievement:

1. what the problem or objective was,
2. what he/she actually did and when,
3. what the outcome or result was,
4. what percentage of the credit he/she claims for the outcome,
5. the name/address/phone number of someone who can verify the achievement.

Once achievements have been collected from an applicant or present employee group, a sample of the achievements is rated on quality by SMEs. Scales containing benchmark achievements -- illustrations of high, average, and low achievements -- are then developed for each dimension on the basis of the extent of SME agreement and scale coverage. These scales are similar to behaviorally anchored rating scales. Typically, achievement scores are combined across dimensions to derive a single score for each applicant.

Unfortunately, one major drawback has been discovered in attempts to operationalize the behavioral consistency method. Substantial portions of applicant populations which complete traditional application forms refuse to complete the behavioral consistency application supplements.

Operationally, these applicants thereby eliminate themselves frcm further consideration in the selection process. Since the limited research evidence available indicates that the extent to which applicants complete behavioral consistency application supplements is unrelated to various indicators of applicant quality. The practice of using willingness to complete this rather laborious application procedure to screen out large numbers o͏f applicants seems arbitrary at best.

This paper describes an exploratory study involving the development and preliminary testing of a selection instrument intended to yield scores similar to those obtained through the behavioral consistency method, but which requires substantially less effort and time on the part of job applicants. If completion of the selection instrument requires substantially less time and effort from job applicants, there is a high probability that significantly larger proportions of job applicant populations will complete the instrument, appreciably enhancing the capacity to apply operationally the theoretical concepts on which the behavioral consistency method is based.

The new selection instrument is called the Activity/Achievement Indicator (A/AI). It consists of sets of three activity/achievement statements. Applicants are asked to choose one statement from each set of three which best represents the type of activity and level of achievement most typical of their respective backgrounds. Each set contains one "high," one "medium" and one "low" activity/achievement statement for a given KSA dimension. The activity/achievement statements are equivalent to scale anchors for benchmark achievement rating scales used in the behavioral consistency method. That is, the activity/achievement statements are derived by means of the same generation, retranslation, and scaling procedures used in the development of the benchmark achievement scale anchors.

Surprisingly, it is not difficult to obtain a sufficient number of activity/achievement statements to create multiple sets or items for each of several KSA dimensions which one desires to measure. When developing behaviorally anchored rating scales, other researchers have found that a larger number of scale anchors survive the retranslation and scaling procedures than are required, thus permitting the development of parallel forms for the performance dimension measures. In the present investigation the survival of a sufficient number of activity/achievement statements would have permitted the development of parallel benchmark achievement rating scales, and did permit the construction of six sets of activity/achievement statements for each of six KSA dimensions.

Method

Subjects

This study involved two different samples. A developmental sample was used to pretest the A/AI (which was subsequently refined on the basis of item analysis), to pretest self-assessment measures for six KSA job dimensions, and to obtain test-retest reliability estimates for the self-assessment measures. The second sample completed the refined A/AI, self-assessments for six KSA dimensions, and a portion of the second sample also completed a job related achievements application supplement used to operationalize the behavioral consistency method.

The developmental sample consisted of 41 college students at a large midwestern university enrolled in a personnel management course during the summer term. Sixteen were seniors; 25 were graduate students seeking Masters degrees. They ranged in age from 20 to 41 years, with a mean of 26.4 and a standard deviation of 6.1. Twenty-nine were male, 12 were female, 35 were white, four were Hispanic, one was an American Indian and one was an Asian/Pacific Islander. One subject neglected to supply sex or ethnic data.

The job-related achievements supplement used to collect behavioral consistency data contains the following instructions:

> In this application form, you are asked to describe what you consider to be your major achievements showing that you have the job-related knowledge, skill or ability identified. These achievements might have resulted through things you have done in any kind of setting--such as paid work, volunteer service, educational endeavors, hobbies, etc. The achievements may be either specific incidents or examples of sustained high performance over a period of time. It is to your advantage to describe two achievements for each skill or ability item contained in this form.

The following six KSA achievement dimensions were derived from a thorough analysis of the job of Product Line Planner:

ACHIEVEMENT ITEMS FOR THE JOB OF PLANNER

ITEM 1. Analytical and Quantitative Reasoning Ability

Planners must analyze a variety of complex information including technical data on product production processes, product cost analysis data, market research data, etc. In doing this they use logic and quantitative reasoning abilities, and must be able to distinguish essential from nonessential information.

ITEM 2. Interpersonal, Organizational, and Coordination Skills

Planners must be able to work with all kinds of people--different socioeconomic and ethnic groups, personalities, age groups, and occupational levels. They must be able to persuade, influence, motivate, organize, and coordinate the activities of people at various occupational levels in several organizational units. Planners coordinate, monitor, and organize the activities of others to achieve certain objectives, but do not have line management authority over the people coordinated. Thus, planners must be sensitive to the needs and requirements of people at different organizational levels, and realize the extent to which they can aggressively promote their own ideas.

ITEM 3. Motivation, Initiative, and Ability to Organize Work

Planners must possess initiative and motivation to learn about new products, marketing and production techniques, sales analysis techniques, and related concepts. They must be able to budget their time for accomplishing tasks and assignments within given time frames and guidelines. How willing are you to seek out and assume additional responsibility and to explore better methods for accomplishing your work? How well can you work with more than one complex project or assignment at a time, organize them as to their relative importance, and allow time for each based on that importance?

ITEM 4. Mechanical/Technical Aptitude

Planners must have the ability to grasp/understand mechanical/technical concepts related to materials and processes utilized in product production. Can you understand basic mechanical/technical concepts related to processes after a brief exposure?

ITEM 5. Oral Communication Ability

Planners must be able to react quickly, confidently, and with composure in stressful interpersonal situations, and present ideas or information in an organized manner on short notice. How successful are you in this type of oral communication?

ITEM 6. Writing Ability

Planners must be able to communicate well in writing. Can you write clearly and consisely?

The achievements recording form is as follows:

Job-Related Achievements Recording Form

REVIEW THE NARATIVE STATEMENT FOR ITEM 1. In the space provided below, describe your achievements which demonstrate your Analytical and Quantitative Reasoning Ability.

ITEM 1. Achievement #1. - - - - Problem or Objective

What you actually did and when (approximate dates):

What the outcome was:

Name, address and phone number of verifying person:

Percent of Credit Claimed: ________%

Benchmark achievement rating scales were used by the T&E evaluators for each achievement dimension.

A detailed account was made of the methodology, the results and an interpretation thereof. The A/AI requires little developmental effort beyond that required to develop a behavioral consistency selection procedure. In terms of total scores, the high correlation (r = .58) between A/AI and behavioral consistency results suggests that the A/AI may have potential as an alternative to the unpopular (with applicants) behavioral approach.

There is substantial room for improvement in the A/AI as operationalized in this study. The internal consistency reliability estimates need to be raised. Individual A/AI items might be improved by making them less subject to distortion. The activity/achievement statements could be degrouped together in fours based on statistical properties of social desirability in addition to performance level discriminability, thereby yielding items more like those found in forced choice performance appraisal scales. The behavioral consistency method can be improved, also.

Both of these procedures have problems. Considering separately the dimensions which these procedures were designed to measure, convergent validity was obtained for oral communication ability and writing ability only. Given the lack of convergent validity for the other four dimensions coupled with the virtual absence of discriminant validity, one is left in nearly total darkness with respect to what it is that is being measured with some consistency by both of these procedures when they are taken in their respective entireties.

Oral Exam Scores: An Introductory Investigation of Minority and Non-Minority Interaction Effects

Sydney L. Teske, Hennepin County Personnel Department, State of Minneapolis

Two data collection efforts were summarized showing some Rater, Ratee interaction effects which can have a significant impact on both race and gender of Ratees.

Method: In 1982 Hennepin County asked the test administration staff to record Race and Sex information for each Rater and Ratee across all oral examinations for about 6 months. In addition, each Rater's score for each candidate was recorded separately. The only other variable of interest was whether the classification tested for was Male or Female Dominated.

Results: Possible bias was indicated in several cases.

1. Female candidates were rated differentially from male candidates when rated by an all male panel.

2. Female candidates for male dominated jobs when rated by male or female raters, received different scores from scores given to male candidates.

3. Male candidates for female dominated jobs were rated low by all raters.

4. Minority candidates for female dominated jobs were rated differently by female raters than male raters.

5. Non-minority candidates for female dominated jobs were rated lower by minority raters than non-minority raters.

The same study was reinitiated in 1984 and data was collected by the City of Milwaukee, the State of Washington, and again by Hennepin County. The same method was employed; however, the purpose was to evaluate whether changes had occurred since 1982.

Results:

Substantial improvements were identified which reduce the disparity noted in 1982. It would appear that the gender and race gap is closing somewhat; however, two points were noted.

1. Minority candidates are scored differently by minority raters than by non-minority raters.

2. There may be some "same sex" bias by female raters.

Comments:

Exam administrators need to collect and evaluate this data in order to determine the impact of the oral exam components from a race and gender perspective. Different cultural environments from jurisdictions may have marked effects on the score results if not monitored and evaluated. In certain, though by no means all in all cases, it may be advisable to balance oral boards with respect to gender and race.

Alternatively, with the knowledge gained from such data collection and analysis projects, other changes can be implemented which also lessen the disparity in rater evaluations. This is especially important for classifications where few or not qualified minorities or females are available to serve as raters on an oral panel.

The Development of Reliable Oral Interview Procedures for Promotion Candidates

John T. Fɪynn, University of Connecticut; Barbara E. Anderson,
James J. Rubovits, Rhode Island College

While there are a number of acceptable ways of estimating the reliability of paper and pencil tests, the focus here will be upon the reliability of one rather troublesome evaluation procedure, the oral interview. The oral interview, which is used with considerable frequency in employment decision-making, is troublesome because the decisions made from it are often influenced by a variety of factors not directly related to the applicant's job suitability. Research has shown that employment decisions can be affected by such nonverbal behavior as the candidate's eye-contact, body language, and gender. It has also been found that the interviewer's non-verbal behavior affects the candidate's interview performance. These kinds of effects can lead to two equally qualified candidates receiving widely discrepant ratings.

The most important step in creating a reliable oral procedure is to standardize both the oral itself and the scoring procedures. In addition to standardizing, i.e., they should be rated on the same scale for all candidates. Since numerical data are required for estimating reliability, the numbers assigned to each candidate's response should have the same meaning for each rater. In order to further preserve the reliability of the scale, multiple raters should be used.

A reliable scale is one which results in agreement among raters for each question and for the total score for each candidate. If the raters, after hearing a candidate's response to a particular question, give widely disparate scores to that response, and continue to do so on other questions, then the scale is unreliable, therefore invalid, and of absolutely no use in making rational, fair, and defensible employment decisions.

The present study describes the development of an oral interview procedure used as part of the police promotion procedure in a small northeastern city.

A job analysis was conducted by surveying all incumbents and immediate supervisors of incumbents in the rank. Job analysis subjects were asked to provide a list of "important" job characteristics. Responses to this survey resulted in a list of 40 characteristics.

The list was then returned to the incumbents and supervisors, who were asked to rate, on two Likert-type scales, how frequently each behavior was performed and how important the behavior was perceived to be in successfully performing the job. Statistical scaling of the responses yielded information on the relative frequency and perceived importance of each of the 40 characteristics. Characteristics were then grouped into categories representing broad classes of behavior, such as verbal expression, logical decision-making, and problem solving ability.

The oral was composed of four open-ended questions suggested by the job analysis data. While there are numerous less definable characteristics which undoubtedly contribute to successful job performance, a defensible procedure would include only those which are directly observable and measurable.

The scale upon which each candidate was evaluated was a 10-point Likert-type scale. Only three verbal anchors appeared on the scale: poor, average, and excellent. The 10-point discrimination allowed the raters to place a candidate's response at ten positions between "poor" and "excellent." Such a graphic rating scale has been shown to produce a more accurate rating of responses than other types, and tends to result in a more reliable procedure.

Four raters from other jurisdictions evaluated each candidate's responses on each question. Raters were instructed to evaluate each response independently of the other raters, thereby eliminating the possiblity of "pseudo-reliability," which results when raters confer and agree upon a single score. When discussion among raters is allowed, the scale will appear to be reliable; however, such a practice amounts to having only one rater evaluating the candidates, and the reliability estimates derived will be meaningless.

Fifty-two police promotion candidates participated in the oral. Total score had a possible range of 0 to 40. Reliability of a scale of this type is estimated by examining the extent to which raters agree in their assessment of the candidates. Several methods investigating agreement among raters were employed. The first, a Friedman Two-Way Analysis of Variance, is designed to assess agreement among the four raters. The Friedman is a technique used for a repeated measurements situation when assumptions of parametric tests are not met. The obtained chi-square (x^2_r) of .0008 indicated considerable agreement among the four raters (df = 3. p = 1.00).

An alpha reliability coefficient was also computed. The alpha coefficient represents an estimate of the internal consistency of the scale. The obtained value of r = .91 ($p < .01$) far exceeds the minimum value required for statistical significance, and indicates a substantial agreement among raters. The third analysis, an intraclass correlation which is a slightly different version of the alpha coefficient, resulted in a coefficient of .72. This coefficient was also statistically significant ($p < .01$) and corroborates the reliability of the procedure. Additional reliability data were provided by a matrix of correlations between all possible pairs of raters. Correlation coefficients ranged from .65 to .81.

Mean correlation between raters was .72, and was determined by transforming each coefficient in the matrix into standard score units.

Summary

Fifty-two police promotion candidates participated in an oral examination, and were independently rated on four structured questions by four raters. Results of statistical analyses indicated a considerable amount of agreement among raters, and supported the reliability of the procedure.

Are All Oral Panels Created Equal? A Study of Differential Validity Across Oral Panels

Bruce W. Davey, Connecticut Department of Administrative Services

Many testing specialists believe that a high level of interrater agreement among panel members in a structured oral examination is practically an assurance of validity. In addition, there seems to be a general belief that a high level of interrater agreement for a structured oral panel indicates that the structured procedures have worked, and that if a second panel followed the same procedures, it would agree closely with the first Research data are presented which disconfirm those beliefs, and remind us once again that while reliability is a necessary condition for validity, it is not a sufficient one. It also reminds us that there is a difference -- and sometimes a major difference -- between intra-panel agreement and agreement across panels.

To overgeneralize the issue, I think there are two major factors contributing to the reliability of the typical oral exam. High reliability, if it exists, is due in part to the structure of the process, and in part to group dynamics. While structure provides the panel with a set of standards to help them form their judgments, I think those standards are fine-tuned considerably through the give-and-take of the panel's members. This would assure a high level of internal consistency within the panel, but would assure nothing across panels.

The present study afforded an opportunity to study this. The study took place in a "live" setting -- a structured oral examination administered to 709 candidates for State Police Trooper, using six separate oral panels. Since there was a great deal of data available concerning each candidate in the process, and since more than 100 candidates were ultimately hired, an opportunity existed to study the construct and predictive validity of the ratings of each of the six oral panels.

PAPER SESSION

Non-Traditional Testing Methods and Uses

Chair: William Tomes, South Carolina State Personnel Division, Columbia, South Carolina

Resolving the Ante-Career Crisis with Military Job Applicants[1]

The opinions expressed herein are those of the author and do not necessarily reflect official Department of the Navy policy.

Herbert G. Baker, Navy Personnel Research and Development Center, San Diego

Generally, family and school have failed to prepare the young man or woman to make a wise occupational decision. Further, in today's occupational structure there is an increasingly broad array of job alternatives that must be considered. As a consequence, entry into the adult world of work is difficult. "Antecareer crisis" denotes the dilemma that confronts the typical American youth seeking entry into his or her first full-time work experience: It is a crisis of unpreparedness. In many selection procedures too much attention is given to efficiency and too little to satisfaction and enjoyment. It is critical to measure more than cognitive aspects alone. Essentially, the military job applicant needs some means of self assessment, adequate occupational information, and a method to link the two: In short, some vocational guidance.

To be sure, there are excellent guidance and counseling systems available in the civilian community, but their recruiting compatibility is marginal at best. Further, we must not expect the recruiting services to enter willy-nilly into vocational guidance. Guidance is neither the mandate nor the primary function of recruiting. What is needed is sound investigatory activity. That is, we need research into the feasibility and the advisability of testing and counseling during the recruiting process; research into effective instrumentation--all of it clearly sponsored by the recruiting agencies themselves as the controlling authority and as a potential beneficiary. The search must lead away from instruments, methods, and procedures that would be inimical to the accomplishment of recruiting's mission--no matter how effective these methods and procedures might be in experimental or educational contexts. Instead, the search must be for recuiting-compatible applicant assessment and occupational opportunities exploration methodologies, that will benefit job applicants and the armed services, while impacting favorably on recruiting operations.

None of these proposals is new; nor is research on interests, preferences, values, and biodata as related to enlistment. There have been some promising starts. In addition to development of instruments, some attempts have been made to design guidance systems, such as the prototype AGENA System (Sands, 1980) for Navy recruiting. To date, no system has had the necessary acceptance. Hence, none has been tested. Moreover, none has had a thorough enough conception to provide adequate guidance through self-assessment and job opportunities exploration. The suggestion is, then, that a recruiting-oriented vocational guidance system must include applicant assessment procedures (aptitudes, skills, interests, preferences, and values--and possibly biodata; plus occupation information tailored in meaningful terms to entry level job applicants--linkable to personal information; and a bridging mechanism to match applicant to job openings.

As the scarcity of applicants increases, there may be a deepening concern for the young man or woman in the crisis of occupational decision. It is readily apparent, in any case, that help in the way of vocational guidance is needed by our youth: Nowhere is this more true than during entry into the very special occupational melieu of the armed services. Assisting individuals to know themselves better is important because persons with inaccurate self-knowledge make inadequate choices more frequently than do those with more accurate self-appraisal (Holland, 1959). Increased match between individual characteristics and the assigned job may be one way to reduce premature attrition and increase performance, job satisfaction, and interest in a military service career.

Forced-Choice Reference Checklists, Adverse Impact and Test Fairness

Barry E. Knake, Puget Sound Naval/Shipyard, Bremerto.., Washington

The contents of this paper are the responsibility of the author and do not necessarily represent the official policy of the U.S. Navy.

Introduction. The objective of Title VII of the Civil Rights Act of 1964 according to the U.S. Supreme Court in their unanimous precedent setting decision (Griggs vs. Duke Power, 1971) was to end discrimination in employment through objective "color blind" measures of ability to perform the work. However, the review of literature to identify such measures by Reilly and Chao (1982) found no valid employment procedures which do not produce adverse effects. This literature review missed the author's 1980 IPMAAC paper on the construction of reference checklists from job element study results and Lilienthal's (1980) review on the use of reference checks for selection. This paper summarizes current research on this content valid selection procedure which is not demonstrating serious adverse impact on women or blacks in nontraditional employment settings (trooper, warehouse, and bus operator).

Validity of forced-choice ratings and utility of reference checklists

The usefulness of forced-choice ratings at distinguishing job performance levels is well documented. (References were cited). Richardson (1949) was first to document the scientific merits of this rating format. Behaviors which are most integral to effective job performance are objectively identified by several master workers following operationally defined procedures. The rater's task is thus simplified to one of describing rather than evaluating the ratee's behavior (by choosing between 2 attractive statements the one statement which best describes the ratee).

The reference checklists and self ratings are extensions of this basic logic to the employee selection or promotion program. Reference checklists must, by necessity, restrict performance domain coverage to those job elements which are rateable in the applicant reference population. These job elements are readily identified by collecting ratings of job element rateability from master worker references, job applicants and/or other sources naive of the job's integral requirements.

Ratings of each element's social desirability are generally collected concurrently from the same population that rateability ratings are collected from and used to help control for the transparency of keyed items on the checklist. This is accomplished by pairing important elements with relatively unimportant statements which are as close as possible in their social desirability.

Relationship of reference checklists to other valid indicators of ability to do the job. Job elements are first scaled according to their criticality to effective job performance; the checklist then pairs critical with noncritical elements, equating each statement in the pair according to their social desirability. The applicant's score is a function of how many times he/she is described by the critical job elements (behaviors).

Research on the relationship between scores on the reference checklist and other valid indicators of job performance ability help confirm or disconfirm the construct validity of reference checklists as measures of ability to do the work. In a previous paper by the author significant, positive correlations were reported between forced-choice reference ratings and self forced-choice ratings, written critical incident test scores, a memory test, a writing skill performance test, a physical ability test, and a background investigation (the structured oral did not correlate significantly). Each measure was developed from a job element study of Washington State troopers. With the exception of the written critical incident test, these correlations were based on checklist scores restricted in range. A study by Dilly also reported positive, significant, and useful correlations between supervisory reference checklist scores and written critical incident test socres for Alaska State troopers.

A study of the interrelationships of reference checklists ratings (by first and second level supervisors), written critical incident test scores, self forced-choice ratings, and interest and willingness checklist scores was conducted on a 1982 administration of a promotion program for pipefitter foremen at the Puget Sound Naval Shipyard. Each of these measures was developed from a job element study of pipefitter foremen. All correlations are in the expected direction and (with the exception of self ratings with second level supervisor ratings) are statistically significant at the .05 point of significance and are useful in magnitude.

Each of these measures sample the performance domain of pipefitter foreman competency. Although there is some content overlap in the subelements sampled, the underlying construct is ability to do the job. The significant relationships achieved between the forced-choice supervisor reference ratings and these other methods for assessing ability to do the job indicate good construct validity for the reference ratings. This is consistent with other correlational studies on the construct validity of forced-choice reference checklists.

Inadequacies of statistical test fairness theories.

Title VII addresses the effects of discrimination against individuals on account of race, sex, color, religion, and national origin. EEO enforcement agencies, as an administrative expedient, have adopted the 4/5th rule of comparing the selection rates for these groups as a means of determining if prima facie discrimination is evident (Uniform Guidelines on Employee Selection Procedures, 1978; however, the courts are not bound to this administrative tool (Connecticut vs. Teal (1982)). And comparisons between groups remains the central theme behind most psychological definitions of test fairness: Thorndike (1971), Darlington (1971), and Cleary (1968).

These approaches to measuring test fairness ignore a critical requirement to the concept of what is fair, that is the fairness of the selection process to all individuals affected. Individuals who have the most ability to contribute to the mission of the work activity should have a higher probability of being selected than those individuals with less of this ability. As all real jobs require a constellation of diverse behaviors for effective mission performance, fairness compells the sampling of all these behaviors in the selection program. If a behavior (or related group of behaviors) is over sampled or weighted in the selection program, the program unfairly discriminates against those individuals who possess the under-sampled or under-weighted behaviors. If the oversampled or overweighted behaviors are related to race, sex, color, religion, or national origin the employer incurs back pay liability and the imposition of quotes even if the validity of its program is well documented if other, less discriminatory selection procedures were available (Albermarle vs. Moody (1975)).

Different individuals get an advantage in employment opportunity depending on the content of the employment practic . Fairness can only be approximated by thoroughly defining the competencies integral to effective job performance and fairly representing these behaviors in the resulting selection program. It is only through valid definition and sampling of all behaviors integral to the job that fair selection practices can be achieved.

A fatal assumption underlying traditional, statistical definitions of test fairness is that the criterion used to judge job performance is a representative and uncontaminated measure of the worker's contribution to the aims of the work. Contaminated ratings can produce spurious statistical correlations, leading to a misleading conclusion that the test measure is even valid. Flanagan (1974) reports research on the interrelationships of Naval Officer performance ratings which indicates that the typical graphic performance appraisals are highly contaminated by the ratee's general reputation. If the tests sample the same behaviors which contaminate the criterion (e.g., intelligence or G factor) erroneous conclusions of fairness (as well as validity and validity generalization) can result.

Criterion sufficiency is also a critical assumption which few criterion measures can satisfy. Assuming the criterion samples behaviors integral to the aims of the work, the representativeness of the criteria is essential if a claim of fairness is to pass scrutiny. If job performance is judged in restrictive ways, the unfairness of the "correlated" test (or content valid measure of a job behavior such as reading comprehension) to individuals who excel at unsampled behaviors integral to the job is obvious.

Employment Test Fairness. Although complete fairness to all job applicants will probably remain an elusive goal, progress in this area is achieveable only through careful attention to the content validity (representativeness) of the behaviors sampled to the total examinable performance domain. Test fairness is achieved when the selection program is a content valid sample of all behaviors integral to effective job performance which are practical to expect from job applicants. The reference checklist is proving itself as a useful tool for approximating this objective.

The relative performance of blacks, whites, men, and women of reference checklists in three nontraditional employment settings was studied. The difference between white and black means is less than 1/2 of the standard deviation for either group in state trooper, warehouse worker, and bus operator employment settings. The results for men and women is similar (except warehouse worker where a small sample of women (n = 12) outscore men by almost one standard deviation).

These results support the contention that fair sampling of a job's examinable performance domain will lead to fairer measurement results, that is, less adverse impact. Reference checklists open the selection process to measures of typical performance on many critical job elements considered "unmeasurable" by traditional credentialism and aptitude/intelligence measurement practices. These job elements include competencies such as honesty for state troopers, ability to stay out of arguments for bus operators and willingness to give a day's work for a day's pay for warehouse workers.

Strange Bedfellows: Work Sample, Content Validity, Trainee Class

Ollie A. Jensen, Educational Testing Service

"Work Sample, Content Validity, and Trainee Class" are strange bedfellows because (1) the Uniform Guidelines on Employee Selection Procedures imply you cannot use a work sample approach to selection for positions in a class in which those appointed learn to do the work after they are hired and (2) the draft Joint Technical Standards for Educational and Psychological Testing imply that only predictive or concurrent evidence of validity can be generalized or transported. Neither of these facts are true.

The author presents an example of transporting a content-validated work-sample test across countries, organizations, classes, occupations and time after it has been established that the content validity strategy may be used to develop selection tests for positions in classes in which the employees are expected to learn to do the work on the job.

The work-sample approach to selection test development is appropriate for any class of positions. If employees must learn to do the work after being hired, the learning-sample form of work-sample test is appropriate.

Both learning and achievement work-sample tests can vary as to directness of measurement. They can vary along a whole-part item-focus or atomization dimension from duty to task, to task element, to facet of task element. They can vary along any of several departicularization or generalization dimensions, e.g., (1) from what is characteristic of one position or a small cluster of positions to that which is characteristic of a large cluster or several clusters of positions, (2) from what is characteristic of the immediate organizational unit to that which is characteristic of an organizational hierarchy or several units or hierarchies, and (3) from what is characteristic of one discipline or field of work to that which is characteristic of several disciplines or fields of work. As the number and kinds of analytical steps increase, the directness-of measurement classification of the appropriate testing instrument changes from that which is most direct to one that is less direct: from "performance of on job site" to "work simulation", to "departicularized work simulation", to "job knowledge or competency test", in which items measure appropriate what, when, where, how, why aspects of departicularized task elements.

To apply the content validity strategy to a job knowledge or competency test in which the coverages are based on the findings from a work sample analysis, there must be a demonstration that each step taken in the atomization and departicularization process logically follows from the previous step. The critical factor is not how indirectly the test measures job performance variables or the width of the inferential river between position performance and test performance; it is the length of the greatest distance between any two stepping stones of content-validity-supporting evidence in the river at the point of crossing.

Due to several circumstances, the work-sampling approach to test specification and development has not been used extensively. The usual approach is to give an examiner and a group of job authorities the results of a job or task analysis and ask them to infer and operationally define the KSA's needed to perform the duties or tasks of interest.

The items in the resulting list are usually a mix of psychological constructs (e.g., verbal comprehension, inductive reasoning, spatial scanning) and task or task element statements with the phrase "Ability to" or "Knowledge of" stuck in front of each one. For instance, "Prepares budget drafts, within stated guidelines and constraints, covering local office operations for submission to central office."

Psychological predictor constructs cannot be supported solely by the content validity strategy. Thus, test specifications that are mixtures of psychological constructs and of paraphrasings of task statements cannot be solely supported by content-validity evidence regardless of how much face validity the operational definitions of those constructs may have. On the other hand, tests which sample those aspects of the work that are important validity evidence regardless of how indirect, abstract, or lacking in face validity those measurements may be.

The first two of the following four examples of how departicularized learning sample tests may be used to select persons for positions in which those selected learn to do the work after they are hired, also show how a test developed under the content-validity strategy may be transported from a class in the uniform service of California state government in the mid 1960's to a trainee clerical classification in the office service of a United Nations agency in Rome, Italy in 1980.

In the mid 1960's I was asked to develop a selection test for toll collectors on the bridges in the San Francisco Bay area. The major problem with the selection procedures used previously was that all those reachable on the elegible lists were students in the local colleges and universities. The two main problems with the students, once hired, were (1) most only worked for the equivalent of one semester and (2) within a week after being hired, most had difficulty maintaining the accuracy requirement of ± 25¢ in each $1,000 in tolls taken in.

Job analysis indicated that the primary task of the toll collector was to make change for 50 minutes at a time under conditions requiring moderate speed and a high degree of accuracy. The change making operation was learned on the job during the first 50-minute work period. Making change is a simple counting operation involving recognition of the amount tendered and then counting from the amount of the toll to the amount tendered using denominations available in the register.

A learning sample test which utilized a counting operation slightly more simple than the one learned on the job was developed. Minimum speed and accuracy requirements, slightly lower than the typical job requirements, were also set. The learning sample consisted of the following: (1) directions for taking the test, (2) an example plus an explanation, (3) a five-minute practice test with feedback as to correct answers, interpretation of speed, accuracy, and number right (useful production) scores and an opportunity to ask questions, and (4) the 50-minute test for the record. The minimum speed score was 100 attempted, the minimum accuracy score was 95% right of those attempted. The rank ordering score was the number right score for those meeting the speed and accuracy cut-offs. Under these conditions, a number right score could vary from 100 to 270.

Fifteen years later I was asked to develop a selection test for trainee clerical positions at the Food Administration Organization headquarters in Rome. The organization hires all clerical employees at the lowest levels and then promotes from within. As there was and is a high unemployment rate in Rome and the organization was and is a preferred employer, many promotable people were and are being hired. These people do well once promoted to higher level clerical jobs or to administrative assistant jobs. At the lower levels, however, many consistently make far too many errors. Many cannot maintain a high level of accuracy on the simple repetitive routines (learned on the job) involved in working as mail clerk, messenger, or document sorter, counter, or filer.

Job analysis indicated that all the routines involved from one to two decision points and 4 to 6 count or identify, match or differentiate, code or mark steps. The speed requirements were moderate; the accuracy requirements were high; the typical between-break work period was 45 minutes.

The test specifications for a learning-sample test (departicularized so as to encompass all benchmark positions in the entry classification) turned out to be the same as the test specifications for the learning sample test for toll collector.

Voila! A content validated test is transported from California to Rome, from a uniform service to an office service, from a toll collector class to a clerical class, and from 1965 to 1980.

A follow-up of toll collector appointments was conducted six months after hiring. Management reported all appointed were on the job and performing satisfactorily.

In Rome, a concurrent study was run in which categorical ratings of job performance were matched against the same three categories of test scores:

1. high accuracy plus well above standard speed

2. at standard to high accuracy and at standard to slightly above speed

3. below standard accuracy and/or speed

The categorical test scores for all 36 employees tested matched their performance ratings.

In both instances there was a negative correlation of about 0.2 between number right test score and amount of formal education completed.

The next and last two examples point up the vagaries of position classification. In both instances there is a journeyman level of job to be done. In both instances recruitment is at the trainee level. In one instance an appointee moves from classroom training to on-the-job training to subjourneyman performance to full journeyman performance within four months and one classification. The "all levels within one class" example involves relatively complex seasonal clerical work. The job is to resolve errors made by persons filling out financial aid forms that are coded on computer-generated correction documents. In the other instance, there are three separate classes (trainee, subjourneyman, journeyman) and it takes 2 to 3 years to reach full journeyman performance. The other example involves building a learning-sample selection test for systems analyst work. Here, a test is used to determine whether a candidate with no systems training and little or no information-processing experience can learn to perform systems analyst work at the full working level (i.e., will be promoted in the normal course of events and within the normal time frame to the journeyman level classification).

Summary. 1. All selection tests developed from properly executed work-sample analyses can be fully and solely supported by evidence of content validity.

2. Every selection test should be supported by as many kinds of validity evidence as feasible within the given set of fixed institutional constraints.

3. A work-sample test or tes item may measure at any level of responsibility from duty to facet of a task element and at any level of position specificity from that subject matter and methodology that is peculiar to a single position in one organization unit to that which is core to many positions in many classifications in many organizations.

4. A work-sample test may measure present ability to perform work activities at a prescribed working level of competence or it may measure ability to learn to perform work assignments at a prescribed working level.

5. It is possible to transport tests developed under a work-sample strategy across organizations, occupations, classes, or time.

PAPER SESSION

Law Enforcement Personnel Selection and Retention

Chair: Samuel J. Bresler, District of Columbia Office of Personnel

Reliability of a Situational Judgment Test for Management

Roger Davis, King County Personnel Department, Sea. ..

Normally, we like to report only our successful experiments; conventional wisdom has it that there is little or nothing to be gained by publishing the results of our failed or only partially successful experiments. But that is not always true, and in this case there are things to be learned from some particular test administrations that were not more than partly successful.

Small group testing presents certain validation problems, the requirements of which are typically met of course through rationales of validity: job analyses and content validity procedures, transportable validity studies, review and approval of subject-matter specialists and master workers. All of these are advance estimates of expected validity, predictions of preditive power as it were, even if the predictions are not expressed in quantified terms. On the other hand retrospective validity estimates, after the small-group test, are another matter, too often ignored or left merely to anecdotal reflection and general approval of the results. While hiring may be very low in small group testing, if the resultant data suggests some of the characteristics of good tests, one can infer through that additional concrete information something more about the value of a particular test. One such characteristic of a good test which is too often ignored is its sheer reliability.

Why all that is of consequence here is that the test in question has (that is, it itself literally does possess) a patent, prescient, and enormous Basis-in-Validity, while in a particular use, or administration, for one client organization, the same test did not clearly manifest that it met a necessary precondition for validity.

To begin with, the science in this paper can be summarized in a single paragraph. King County administered a well-developed, well-known, middle-management test in 1980 and in 1983 to two small groups of candidates. The administrations were completed without incident. In our two administrations, we had 37 supervisors who took this test for management both times that it was administered. Aside from estimates of test-retest reliability, the <u>actual</u> test-retest reliability for the 47 management candidates was low, only a Pearson product moment correlation of 4 = .44.

As we all know, testing for management positions is not currently in great favor, particularly among managers. With the exception of assessment centers and some medical and psychological screening, management testing is not in vogue today, and it is my perception that that is because managers themselves do not favor testing for the managerial class, even while there is an enormous pro-testing wave sweeping over us in this country in the last quarter of the twentieth century.

In the civil service industry, however, management testing is an important responsibility, and in testing for police command management one instrument available to meet this challenge is the Police Career Index (PCI), authored by Dunnette, Motowildo, et al (1976). Through special arrangement with the principal researcher, Professor Dunnette, I have used a sub-test of the PCI in recent years for evaluating King County, Washington police command management candidates. The sub-test is the Situational Judgment Inventory for Intermediate Police Commanders.

As you probably know, situational judgment inventories are based upon the critical incident techniques principally developed by John Flanagan, (Measuring Human Performance, 1962). These tests are typically but not necessarily paper-and-pencil tests. The problems are often arranged in the format of traditional multiple-choice tests.

Two important differences from multiple-choice tests lie in the nature of the problems and in the scoring design found in situational inventories. Instead of being factual questions of knowledge, the contents of these tests become situational problems of judgment. This characteristic often gives the impression to candidates that their test is a compilation of historically true episodes, and consequently gives a powerful sensation of the test's face validity. One such test that I developed for the rank of police sergeant in 1977 was highly praised by both the Seattle Police Chief and the Seattle Police union, but I suspect it was this "war-story" flavor of the test rather than its structural merits which persuaded them my situational inventory was a good test for first-level management. The other characteristic, the scoring design, is interesting because this type of testing allows at least two "right" answers to the question--the best and the worst solutions to each problem. The response alternatives are weighted variously instead of dichotomously, and candidates earn credit on each problem, according to the stipulated worthiness of their choices as Best/Worst solutions. This scoring design adds economy to a test because it means that an equivalent number of situational inventory problems will have more spreading power than a traditional multiple-choice test of knowledge is likely to have, so long as the response alternatives' weights are "true" weights, i.e., their values are derived from job experts or subject-matter specialists.

Tests such as this, if they are of any length at all, I believe always have true content validity, but I am not, however, really going to join here an argument over content validity as we know and talk about it today, and whether it is a really adequate criterion for evaluating this type of testing.

Perhaps a better question is whether or not such a test contains within it any extraneous material, anything which is job-irrelevant, as it were. Developers and users of such tests, particularly when such tests are intended for broader than localized, specific use, need to be sure that the contents of such tests are within the domains of the target jobs locally. Now that to you may suffice as a pragmatic version of content validity. As a practitioner in the employment arena, the exclusion of extraneous contents is essential in my tests.

One efficient technique for accomplishing the localizing of such test contents is to have the appointing authority assign a panel of job or content experts to take the test. The responses of the content experts can be analyzed to produce a model key based on the consistency, high agreement, or "reliability" of their responses. This localized key effectively edits out the scoring of the test any extraneous content. The test is also factor job-related. More than that, it is job relevant. This step does at least two things additionally: it assures that the judgmental problem-solving process in the test represents the best available thinking of a group of local managers. There is nothing idiosyncratic in the test. And, although this test process is not necessarily predicated on local job analysis or content validity per se, it does assure that the complex scaling and weighting of the response solutions in the test derives directly from a panel of specialists or expert workers. This is the procedure that was invoked by Professor Dunnette so that his nationally standardized test would be locally specific, as in King County.

The descriptive statistics for the two administrations three years apart were quite similar.

TABLE I

King County Administrations

Situational Judgment Inventory for Police Commanders

	1980		1983
N =	61	N =	54
Range	67.5 lo - 88.6 hi	Range	69 lo - 88 hi
$\bar{x}$ =	80.1	$\bar{x}$ =	79.5
SD =	4.7	SD =	4.1

It appeared that everything was almost the same in the two test administrations. Unfortunately, the actual retest reliability was low.

The candidate populations were 61 in 1980, and 54 in 1983. Among the latter 54 were some 37 people who in effect were being retested after three years. Unfortunately, among those 37 the retest reliability was r = .44.

A study of changes in scores showed only 14 persons increased their scores over 1980. Eighteen candidates went down. Only 10 of the 37 retested earned a score that deviated not more than ± 1 point from their 1980 scores. Almost half the candidates earned retest scores ± points from their first-time score.

Why would such an apparently broadly useful test, a test that appears to be an inherently good test, turn out to be not highly reliable, and therefore not highly valid, in this organizational context?

Further analysis of the results suggests that while some candidates rigidly stuck to their answer pattern of three years before--and they likely could remember with vivid clarity how they had responded to these problems--it turned out that many candidates revamped their problem-solving strategy drastically. The seven greatest changes in earned scores were from sergeants who had not been successful enough in the 1980 administration to earn a promotion. But not all these drastic changes were for the better. The greatest change was by one individual who went up 11 points. The next greatest changes were by two persons who each went down 9 points from 1980 to 1983. The third greatest change was by another person who lost 7 points on the re-test. This instability in candidate judgment affected the statistical reliability of the test, and therefore its validity.

Nor were the stable candidates, in terms of the re-test, by any means competitively better than their peers. None of the seven whose scores varied ± point across time were competitive enough on this test to earn a promotic

Instead the candidates who did best in competitive terms were the ones who by and large stuck with their same essential prior response patterns yet tried to make just a few adjustments in their problem-solving, as if time and experience had allowed them to fine tune their situational judgments.

At the beginning of this paper I referred to the topic as a less than successful "experiment." In personnel assessment, however, we work in a non-experimental envirnment. Almost everything we do is final, and we live with it, as occurred in this case, irrespective of which set of scores was more truly valid.

As for the test itself, the Situational Judgment Inventory for Intermediate Police Commanders is a professionally well-developed management tests, unarguably high in its Basis-in-Validity (properties giving indication of probable validity). But in its administrations in metropolitan Seattle it was found to have insufficient retest reliability, perhaps more due to candidate test-taking strategy than to the intrinsic properties of the test.

The Use of a Field Tactical Simulator For Selecting First-Line Police Supervisors

Patrick T. Maher, Personnel and Organization Development Consultants, Inc., LaPalma, California

First-line supervisors in police agencies have as a major work behavior the requirement that they respond to tactical situations, such as a barricaded suspect, major crime scene, disaster, officer involved shooting and other situations, assume command and coordinate the responses of various resources. Such situations require the supervisor to make many critical decisions in a short period of time, under rather stressing conditions. As critical as these situations are, many police executives and promotional candidates believe that current assessment procedures for promoting and selecting first-line supervisors do not adequately test for the abilities acquired of the work behavior. Traditionally, examinations that may touch upon these incidents are limited to paper-and-pencil tests that measure a specific knowledge - oral examinations that ask candidates to describe what they do, performance evaluations that attempt to assume what the candidate might do if confronted with such a situation, and sometimes a planning problem made part of an assessment center exercise. None of these assessment procedures, however, directly assess how a candidate reacts when required to take command at a tactical operation. Even assessment centers are limited to testing generic management abilities and skills that are not directly related to this work behavior.

When recently confronted with developing an entire assessment process for police sergeants in which the ability to handle tactical situations was identified as a critical area that should be measured, it was decided to adapt a simulation process as part of the promotional assessment procedure to measure such abilities.

The score achieved on the simulator was weighted at 30% of the final score on the examination. This weight was determined by the job analysis, which considered the number of KSAPs being measured. The simulator was developed in conjunction with a committee composed of officers from the department and a sergeant representing patrol, investigation, and administrative functions. The committee's initial role was to design the type and scope of the problem to be presented. They were responsible for creating a problem that was as realistic and accurate as possible. However, as with any simulator, a problem was that of determining the correct response to a given situation. This is not difficult when procedure is identified in a manual or other directive. However, much of what is done in a tactical operation is purely judgemental. In these cases, a clear consensus must be developed from the committee and it should be written down, reviewed, and signed by each member of the committee.

The simulator serves two purposes. It measures some technical knowledges and principles of police tactical operations, necessitating that the raters have technical background in law enforcement. The simulator also measures specific knowledges and practices of a particular department. Raters, therefore, must be trained in specific responses of the department involved in the tests, as well as the testing procedures itself. We have used--as raters--individuals who have been trained in and who have experience in the assessment center method. Thus, they have extensive experience in recording behavior and using this rating format.

An important aspect of administration is to insure that candidates fully understand what is expected of them. This is accomplished by leading them through the process step by step, giving them instructions in as simple a format as possible, and giving them instructions orally and in writing. Candidates are next given information concerning the situation that they are going to encounter.

The simulator described here requires the candidates to actually commit resources and take action, or make decisions based on the department involved. A timed script is used to give a proper arrival time, sequence of arrivals resources, or occurrence of activities and to provide other information. All time is in exercise time, but it frequently parallels real time. The time required for this portion of the simulator is dependent on the scope of the problem presented, but requires at least 10 minutes. When the problem is completed, a question and answer period is used by the raters to clarify any questions they may have about why the candidate took an action that may have seemed inappropriate. It also allows them to clarify exactly what the candidate did if confusion arises.

The scoring of the evaluation involves a two-part process: 1) the development of a structured rating form that relates to specific procedures and areas identified by the expert committee and 2) the actual assigning of a score itself. The rating form covers the major areas that the raters should consider. All rating factors that relate to a specific operation or procedure should be referred to a specific manual page, order, training bulletin, committee consensus record, or other document. After all questions about a candidate's performance have been clarified, the raters assign scores. The assignment of scores is done separately from the observation of the simulator performance. Raters also assign scores independently of one another and after all participants have gone through the simulator, scores are assigned on a 1 to 5 scale in each category by each assessor. Once they have scored all candidates in each category, they meet as a group to poll information on each candidate.

Scoring was designed to reflect distinct levels of performance that can be more reflective of differences in school grades. For instance the difference between a 5 and a 4 is similar to the difference between an A and a B. Scores values were converted into an overall percentage score. The results are a more objective process with information that can be articulated to both the candidate and any reviewing authority. Scores obtained using this method are well distributed, do differentiate between candidates and are comparable with Civil Service systems requiring ranking of candidates on a scale of 100%.

Psychology and Law Enforcement: Hiring the "Right Stuff" and Keeping Them Effective

Harold Brull, Personnel Decision, Inc., State of Minneapolis

Based on an extensive research study conducted for the Nuclear Regulatory Commission concerning behavioral reliability in nuclear plants, the author presented a three-part approach to psychological screening and maintenance of a trustworthy, reliable, and emotionally stable workforce.

The screening component for this program includes psychological testing and follow-up interviews when test results indicate the need. The presentation will describe the tests used, the decision rules applied, the interview structure, and the final screening decision sequence. Attention will be paid to defensible use of psychological tests, fairness to applicants, and adverse impact data.

The second phase is a program to equip supervisors and command personnel with the skills necessary to ensure continued reliability on the part of personnel under their direction. The program takes the form of a two-day workshop which can be shortened or modified to complement existing supervisory and management training programs. The program content includes the following areas:

- Models of behavior, people, and change
- Patterns of behavior change in the workforce
- Practice in behavioral observation
- Stress: Causes, signs, and coping strategies
- Giving feedback on performance
- Administrative procedures, including connection with existing Employee Assistance programs
- Practice in applying skills to actual cases

A major focus of the behavioral reliability program is the distinction between the role of the supervisor and the psychologist. Supervisors and managers play a vital role in the continuation of an effective workforce. By exercising their supervisory responsibility, they complement the role of the mental health professional and the screening psychologist.

This behavioral reliability program can be delivered either by outside professionals or in-house trainers using a leader's guide.

The final component of the program involves a re-screening procedure by which a determination can be made whether to return an officer to the line of duty after an incident or behaviors which call his/her fitness for duty into question. This section of the presentation will include discussion of legal antecedents regarding removal of an officer from service based upon psychological considerations.

Handout materials include very helpful materials form the author's two-day workshop for supervisors and command personnel on ensuring continued behavior reliability on the part of personnel under their direction.

The Development of an Interactive, Multiple Component Police Detective Assessment Process

Joan G. Weiss, Office of Personnel, Washington, D.C.

Detectives in the Washington, D.C. Metropolitan Police Department have a two-level career ladder--Detective Grade Two and Detective Grade One. The Grade One position was established to provide a means for recognizing the special investigative skills of Detectives and for rewarding excellence. There are approximately 450 Detectives on the force and only 25 have been awarded the Grade One Classification.

The Metropolitan Police Department requested that an assessment be developed which would ensure the selection of the best qualified individuals to Detective One positions. The goal of the examining project was to design as comprehensive and cost-effective process as possible for identifying superior detectives.

The first step was to conduct a job analysis. A task/KSA inventory was used. This is cost-effective, yields data which can be statistically analyzed, and documents how examining procedures sample the KSAs associated with task performance. The tasks and KSAs were obtained through on-the-job observations, interviews, and brainstorming sessions with incumbents and first-level supervisors. The KSAs and tasks for Detective One were rated by the same people on importance to job, frequency of performance (tasks only), necessity at entry, ability to differentiate among levels of performance, and which KSAs were necessary for the performance of each task. After analysis of the ratings, 16 tasks and 24 KSAs remained. Weights for the KSAs were derived from this analysis also.

The two major objectives for the examining system were (1) an efficient and effective system for processing a large number of candidates and (2) situation-specific behavioral manifestations of KSAs had to be assessed. The final system was a written examination made up of three components, two job simulation exercises, and a suitability rating.

The first phase was a 54-item multiple-choice written examination with questions about factual information in pursuing investigations; two restricted response essay items which tested arrest, detention and case preparation procedures; and finally the candidate had to organize a "case jacket" of a partially-completed investigation, plan the investigation, and describe it in outline form. These three steps had a 3 1/2 hour time limit. After this first phase, the number of individuals who continue is based on the administrative needs of the Department, the availability of assessment resources, adverse impact, and the psychometric properties of the examining components.

The second phase is a crime scene simulation exercise with an arrest report exercise and a suitability rating. The crime scene simulation requires a candidate to respond to a scene of a burglary, manage the scene, take statements, and direct preservation of evidence. After the role-playing, the candidate uses appropriate laboratory and investigative reports and prepares a report on the pursuit of the case. -The candidate also has to prepare necessary documentation for an arrest warrant affidavit on a separate case.

For the second phase suitability ratings are obtained from supervisory officials. They rate the candidates on the basis of a 5-point behaviorally-anchored scale. The 30-45 minute simulation exercise requires an administrator, three role players, and a three member assessor team. The assessors observe and record on the basis of the behaviorally-anchored scale.

The results of how this assessment process may have worked (if implemented) are not available. The project has been delayed due to a pending court challenge by a number of Detective Two's who were denied promotion to Detective One more than six years ago.

SYMPOSIUM

Alternative Methods of Presenting Questions and Related Information To Candidates in an Oral Examination Setting

Chair: Sidney L. Teske, Hennepin County Personnel Department, Minneapolis
Participants: Sally A. McAttee, City of Milwaukee Personnel Department;
D.J. Patton, State of Washington Personnel Department; and
Sidney L. Teske, Hennepin County Personnel Department

Introduction:

A review of the interview literature confirmed that most of the work has concentrated on administrative aspects of the oral interview and very little work has been done on the content or presentation of oral exam items. It is a thesis of the individuals involved in this study that manipulation of rating scales, amount of structure and the quantity and quality of rater training provided, while important, they do not satisfy the fundamental ingredient of the oral exam process. The content and presentation method of the items have received little attention and we believe are critical if validity is to be achieved.

In spite of many negative research findings, the more recent research has suggested positive outcomes under the following conditions:

a. Job information available to oral board

b. Structured rather than unstructured questions used

c. Behaviorally anchored rating scales rather than general trait descriptions used

d. Rater training provided

However, high interrater reliability does not necessarily mean high validity. Trained oral panels with high interrater reliability differed on validity. Also, some research has found that both structured and unstructured interviews had low reliability and validity.

Method

The three jurisdictions involved in the present report have developed a method of oral exam item presentation which allows for a dramatic change in what material can be covered with little or no added administrative cost.

Under one method, items are given to candidates prior to the actual examining time which permits the introduction of two new variables. First, the questions (stimuli) can be much longer than if orally communicated. Second, the questions can be much more complicated.

including items used in problem solving portions of Assessment Centers, which are multidimensional and which provide more focusing information to the candidates.

The introduction of more complex items led to an additional area not fully researched. It has been quite common for some raters to ask all the questions and to rate the responses (score the exam) after the candidate has left.

Use of complex, perhaps multidimensional items, led us to use one of two scoring methods. On the one hand, if the questions are unidimensional, scores or ratings can be assigned immediately after each response is given. In the other case, interim judgments on dimensions can be made in the same way that Assessors in Assessment Centers keep track of behavior observed in the "center" settings.

For the purpose of our research, we concentrated on two variables.

A. Two types of rating methods were compared:

1. Dimension - Job Dimensions or KSA's are identified as the factors to be rated (e.g., oral communication skill). The response to one or more questions contributes to the rating of each dimension. Each question (stimuli) may relate to one or more dimension.

2. Question - The response to each question is rated separately without reference to job dimensions or KSAs

B. Two types of question presentation were compared:

1. Pre-exposed: Candidates were given copies of the questions prior to the oral exam.*

2. Non pre-exposed: Candidates were not given copies of the questions prior to the oral exam.

The design was a 2x2 unbalanced design with two levels of item exposure crossed with two rating methods. Hence, the four cells consisted of the following pairs.

Pre-exposed items rated by dimension
Pre-exposed items rated by question
Non pre-exposed items rated by dimension
Non pre-exposed items rated by question

The class tested was not held constant. Therefore, each cell contained many examinations for different classifications. An observation was considered to be one oral exam panel for a single classification. No randomization was done, rather simple field data was collected and evaluated. As a result, one cell contained only a single observation.

Hypothesis

For purposes of our preliminary research we sought to investigate only the reliability of the oral exam. The study included three hypotheses, one for each main effect and one for the interaction.

a. Scoring by question will produce greater interrater reliability than scoring by dimension.

b. Use of pre-exposed items will produce greater interrater reliability than use of non pre-exposed items.

c. The interaction effect will be significant.

Results

Intraclass correlations were computed for each cell for each panel, such that the measurement of interest was the intraclass correlations. This statistic was used to estimate the interrater reliability for each oral exam.

The results are as follows:

	N	Unbiased Estimate
Pre-exposed by dimension	1	r = .963
Pre-exposed by question	10	r = .889
Non pre-exposed by dimension	13	r = .904
Non pre-exposed by question	26	r = .882

*Also included as pre-exposed items were in-baskets and written problems completed by the candidates prior to the oral exam and rated by the oral exam panel.

The results did not show difference in reliabilities. That is, none of the three hypotheses were proven to be true. However, the reliabilities are higher than could be expected from information in the literature which describes rater consistency or reliability in the interview.

Discussion

The present study contains some potentially large error variance components including the non-randomization of subjects to treatments, and not requiring the class, test or the raters to remain constant.

Further, while the results are informative in that the reliabilities are very high, there is no validity evidence. It is the intent of the three participants to conduct such research during the next year.

A major difficulty in designing a research model has been the applied settings in which we work. Some variables are very difficult to manipulate without major ethical and legal conflict. For example, in trying to vary the type of exam given or the rating method over a single examination for a classification would expose an agency to charges of unfairness or worse.

PAPER SESSION

Improving the Organization: Innovations in Personnel Administration

Chair: Wiley R. Boyles, Auburn University at Montgomery, Alabama

Measuring Behavioral Effectiveness on the Job

J. Ernie Long, U.S. Office of Personnel Management, Seattle Washington

Background

This presentation is about an approach of evaluating employee performance that we are using in some agencies in the federal government. I developed this particular procedure in 1978 although many of you will recognize the basic measurement structure as what is known as behavioral observation scaling (BOS) as outlined by Gary Latham and Kenneth Wexley in their 1981 book, Improving Organizational Productivity Through Performance Appraisal. I called my version of BOS the "behavioral effectiveness" (BE) approach.

We have come to expect a great deal from our performance appraisal systems. In the federal sector, they must be capable of withstanding legal challenge as well as fulfilling a variety of "positive" functions. To ask that a single measurement tool serve all of those functions, such as providing the basis for terminating someone while serving as an enhancer of communications or a motivator to higher levels of performance, may be stretching the tool beyond its capability. But I believe that the BE approach comes closer than any I have seen to fulfilling such a diverse set of expectations, at least in the context of the federal performance measurement system.

This presentation focuses on the basic structure of the BE approach. This approach is being used in several federal agencies in the Northwest and elsewhere in the country and in several state and local governments.

Structure of BE Performance Standards

Since the Civil Service Reform Act of 1978, the federal government has been required to evaluate employee performance using what are known as "performance standards." These are written standards of performance for what CSRA called "job elements." (The "job elements" are usually job functions--five to eight statements which describe essentially all the work being performed.) In a BE standard, performance indicators are used to describe the qualitative, quantitative, and other standards of performance on each of the elements. There can be as many performance indicators as are necessary to communicate performance expectations in relation to the element. Between 5 and 15 performance indicators is normal. Each performance indicator is then measured on a frequency of occurrence scale anchored at the ends by Always and Never.

The most fundamental concept of underlying the BE approach is that if a person usually does what is expected of him/her, his/her performance element of the job will be considered fully satisfactory. So performance indicators describe the behavior and results that are desired, and hopefully agreed upon, by the supervisor and employee.

We are also required to give a single overall rating on each of the 5-8 elements. This is accomplished by means of well defined behavioral guidelines for the performances evaluated. For the Fully Successful level, these guidelines reflect BE's underlying concept that if you usually do most of the things you are supposed to, as stated in the performance indicators, your performance will be considered fully satisfactory.

Incidentally, I do not recommend use of the adjectives in the overall element rating scale (i.e., fully successful, exceeds fully successful, and minimally successful). In a study I did a few years ago called "Scale Values for Evaluative Words" that was reported at the 1983 IPMA Assessment Council Conference, I found that these were among the most ambiguous scale anchors that it was possible to use. We use them because we are required to--you can use better ones. On the other hand, the scale properties of the frequency of occurrence scale (the Always-Never scale) are good and we have found this particular frequency scale both psychometrically and practically meaningful.

The original BOS system and some users of the BE approach have used a more mechanical system, such as a point system, for summing the Always-Never scale ratings and determining the overall element rating. Such an approach has value in some applications but I believe it tends to unnecessarily complicate the system and imply a promise of objectivity and precision that I do not believe can be fulfilled by any performance appraisal system that I am aware of, including BE. Such systems are useful but I believe should be deemphasized in favor of reasoned judgment in assigning the overall element rating. Our experience has been that "the numbers" don't always add up to a correct assessment of an employee's performance.

Some applications of the BE approach also use a single summary adjective rating--a summary of all the element ratings. This is also possible using a scheme similar to the point system suggested by Latham and Wexley. Such is not a part of the basic BE approach but it does not detract from the value of the BE approach if this feature is added on.

In the BE approach, performance tracking is done by exception. If necessary, words like Usually can be given more specific definitions. In the original BOS, these scale points were defined by specific percentage values as shown below:

Almost Never 0 1 2 3 4 Almost Always

where: 0 means 0-64% of the time
1 means 65-74% of the time
2 means 75-84% of the time
3 means 85-94% of the time
4 means 95-100% of the time

I have favored verbal scale anchors (such as Often Does Not, Usually, etc.) over percentages or other numerical anchors partially in order to counter what many people believe to be a tendency toward "bean counting" (non-meaningful quantification) in the performance measurement process in the federal sector. Beyond the meaningfulness issue, there is also some question about the reliability with which raters can differentiate between percentage values, although Latham and Wexley (1981) report adequate reliability to such ratings even in the absence of rater training. In any case, as noted earlier, the verbal scale anchors in the BE approach can still be given more specific numerical definitions if that is necessary.

Positive Features of the BE Approach

The value of the BE approach is best seen when contrasted with performance measurement methods now in use in the federal government, but it also has several positive features when viewed just by itself. The main ones are:

--objectivity

--accuracy of measurement

--ease of development

--value in communicating performance expectations and feedback

This last advantage, the communication feature, is probably the most important. The purpose of doing performance measurement should be to improve the performance of the individual, and as a result, of the organization. The foundation for that improvement must be the communication of meaningful feedback between a supervisor and an employee. Users of the BE approach comment that because of the ease of communicating performance expectations and feedback, they are able to cover <u>all</u> aspects of the job, including the behavioral ones (which are so often ignored because of the scarcity of appropriate measurement tools) and the quantitative ones (which have been excessively favored just because they are easier to measure and give the appearance of objectivity.)

We have found that BE standards provide a way to measure even the softer (behavioral) aspects of job performance accurately and objectively. Our experience with it has been good and thus far the only corrective actions taken on the basis of BE-type standards have been upheld in the federal appeals process.

Note: The author provided handouts which described procedures for developing performance standards and samples of performance standards for clerk-typist, GS-5.

Task, KSA, Job Demand Factor Ratings as Predictors of Occupational Stressors and Strains

Kevin G. Love, Department of Psychology, Central Michigan University

Traditional stress management approach. Unfortunately, current stress management techniques have not taken notice of the significant stressor-strain relationships. Ranging from exercise programs to biofeedback training to progressive relaxation techniques, the focus of typical stress management procedures has been to teach the individual worker how to cope with the strains being experienced. The primary goal of these methods has been to reduce the individual strains being experienced.

However, without empirical data documenting the effectiveness of stress management approaches the choice and implementation of such a procedure cannot be effective. Perhaps of greater importance is that without a focus on reducing the casual factors (high stressor levels) the coping strategies of stress management programs are at best temporary treatments. An individual employee who loses effectiveness in coping will still be affected by the ever-present high stressor levels which will reinstate previous experienced high strain levels.

Job Component - Stress Linkages

Failure to incorporate in stress management. Perhaps one of the major reasons that stress management programs have not incorporated stressor-reduction components has been the neglect of stress research studies to pinpoint facets of the organization or job in question which lend themselves to program inclusion (i.e., control or manipulation). Stressor measures typically have been construct oriented (e.g., role ambiguity, role conflict, etc.) rather than job component specific. It would be hard developing a program to decrease role ambiguity, for example, without information as to which aspects of the job or organization are the "key" factors which have produced high stressor levels.

The closest attempt to link job characteristics with psychological outcomes such as stressor-strain measures has been the work of Hackman and Oldham (1976). Whereas Hackman and Oldham (1976) were primarily concerned with utilizing their model for job enrichment purposes, it is proposed that their approach, (the linking of worker perceptions and psychological states), can be a useful one for stress management program development.

Linking job analytic perceptions and stressor measures. Few job analysis applications, however, have attempted to incorporate resultant worker perceptions beyond their immediate utility in personnel system development. Based on the demonstration of the worker perception-psychological state linkages of Hackman and Oldham (1976), it is proposed that this paradigm may be used to link worker perceptions as measured via job analysis to the psychological states of job stressor levels.

The present study sought to measure (1) tasks (or job duties), (2) knowledge, skill, ability requisites, and (3) job demands and link these perceptions with measures of four psychosocial stressor levels (i.e. role ambiguity, external role conflict, quantitative work overload and responsibility for people.) Significant linkages among important job characteristics and stressor levels would provide an "empirical definition" of job stressors (ala Hackman and Oldham) which would provide the basis for data-based stress management programs geared at both stressor and strain reduction.

Method

Subjects

The study involved job analysis and stressor measures gathered from 378 factory supervisors employed within a high technology production organization. The typical factory supervisor was a white male between 45-54 years of age. Having attained a high school degree and accumulation of more than 21 y:ars with the organization were the most common characteristics. The typical supervisor worked first shift within a production/assembly department.

Procedure

Job analysis methodology. In order to document the important facets of the job of factory supervisor within the subject organization, a task-based job analysis was completed (see McCormick, 1976).

Stressor measurement. As the final section of the job analysis questionnaire, factory supervisors rated the occurrence of four stressors. The stressor items were adapted from Quinn, Seashore, Kahn, Mangione, Campbell, Staines and McCullough (1971).

(1) role ambiguity -- defined as the extent to which role incumbents understand their job duties, rights, and responsibilities. Two items measured this construct as follows: "How much of the time are your work objectives well defined?" and "How often are you clear about what others expect of you on the job?" Possible responses included: 1-never; 2-occasionally; 3-sometimes; 4-fairly often; 5-often.

(2) external role conflict -- defined as the degree of incongr.:ity or incompatibility of others' expectations. Two items were used, "How often do persons equal in rank and authority over you ask you to do things which conflict?" and "People in a good position to see if you do what they ask give you things to do which conflict with one another." Possible responses included: 1-rarely or never; 2-sometimes; 3-fairly often; 4-very often.

(3) quantitative role overload -- seen as having more work than can be accomplished within a given time period using available resources Four items were used, "How often does your job leave you with little time to get things done?", "How much workload do you have?", "What quantity of work do others expect you to do?" Possible responses for the first item included: 1-rarely; 2-occasionally; 3-sometimes; 4-fairly often; 5-very often. Possible responses for the second through fourth items included: 1-hardly any; 2-a little; 3-some; 4-a lot; 5-a great deal.

(4) responsibility for people -- described as having control over the welfare of others, notably subordinates. Two items were used "How much responsibility do you have for the morale of others?" and "How much responsibility do you have for the health and safety of others?" Possible responses included: 1-very little; 2-a little; 3-some; 4-a lot; 5-a great deal.

These stressors have been shown previously to be related to several strains (see Beehr and Newman, 1978). A composite score was calculated for each stressor, averaging ratings across appropriate questionnaire items. The average stressor levels for factory supervisors were beyond the midpoint of each scale.

Results: Linear Regression Analyses

The four stressors, as measured in the present study, were found to be independent. Linear regression analyses were computed to investigate the relationship among important job characteristics, (i.e., task dimensions, KSA dimensions, and job demand dimensions) and stressor levels (i.e., role ambiguity, external role conflict, quantitative role overload and responsibility for people. For each stressor (dependent variable) three separate linear regression analyses were performed on the data using task dimension ratings, KSA dimension ratings, and job demand dimension ratings as separate sets of predictors (independent variables). Simultaneous entering of all predictors was employed for all of the 12 linear regression equations calculated. Consistent with the job characteristics model of Hackman and Oldham (1976) perceptions of specific job characteristics were linked to high stressor levels, across all four stressors.

Stressor-Job Characteristic Relationships

Role ambiguity and external role conflict were most frequently found to have a significant relationship with specific job characteristics. There was little overlap in the job characteristics significantly related to either role ambiguity or role conflict. These findings were consistent with the perspective provided by Nicholson and Goh (1983) in describing these stressors. That is, external role conflict and role ambiguity have substantially different implications for relationships with job analysis perceptions. Whereas external role conflict is interpreted as an incompatibility among resources, policies, or people, role ambiguity involves uncertainty and lack of clarity regarding role requirements for the individual employee.

Role ambiguity. Those job characteristics significantly related to role ambiguity were reject/defect operations, maintaining personal expertise, team activities, documentation of worker problems, labor relations knowledge, packing and shipping knowledge, product knowledge, solid state knowledge, and the physical requirements of the work. Job characteristics most often significantly related to high levels of role ambiguity involved knowledge needed by the incumbents to perform within their required role. In addition, several related, job characteristics involved potential lack of clarity regarding organizational procedures (i.e., documentation of worker problems, reject/defect operations).

External Role conflict. High levels of external role conflict were significantly related to job characteristics involving direct confrontation among people and/or procedures (see Table 8). Specifically, external role conflict was significantly related to the pressure and pace of work activity, salvage/scrap operations, interface with purchasing, obtaining maintenance for the department, scheduling operations, overseeing production, employee counseling, maintenance knowledge, and making adjustments to personal life. Most of these related job characterisitics involved interfacing with other people within or outside of the supervisor's department.

Quantitative role overload. Quantity of role overload as seen by the factory supervisors, was related to the single job demand of pressure and pace of work activity. Defined as too many things to be done, with not enough time for completion, this relationship was almost intuitive.

Responsibility for people. There was a degree of overlap among job characteristics seen as indicative of role ambiguity and responsibility for people. Most of the job characteristics significantly related to high levels of responsibility for people involved areas of the job of factory supervisor which had direct bearing on employee welfare. The pressure and pace of work activity, budget operations, team activities, packing and shipping knowledge, the physical requirements of the work, and employee administrative activities were related to levels of responsibility for people. It is interesting to note the importance placed on financial duties which relate to high levels of responsibility for people.

Using the findings of the present study regarding factory supervisor stressor levels the following recommendations can be made for specific stressor reduction.

(1) knowledge based training program development should center around labor relations, production, product specifications, accounting, maintenance, and counseling areas;

(2) job engineering should focus on an investigation of current budget operations, defective part procedures, and scheduling of operations;

(3) organization development interventions, such as process consultation, should focus on the conflict among production, maintenance, and packing departments.

Impact for Using Job Analytic Perceptions in Stress Research

Data-based stress management. Reduction of the causal factors of job stress for factory supervisors is more cost effective than continual alleviation of strains through teaching of individualized coping strategies. (i.e. the traditional clinical or medical model of individual health treatment). The methodology employed in the present study goes beyond individual diagnosis and analysis to group measures. While the individual is not ignored, the major thrust is the identification of job perceptions and stressor levels for the majority of job incumbents.

The clinical or medical model of health treatment usually takes the approach of diagnosing the individual and fitting that person into the appropriate treatment mode. Taking the job duties, organizational structure, and business realities as given, the individual would be changed to fit within the existant environment. The job characteristic - psychological state model would allow, on the other hand, a tailoring of the work environment for the "typical" employee.

Regardless of one's orientation, the data indicate that specific job characteristics seen as important in the job of factory supervisor are linked to levels of various stressors. To maximize reduction of stress levels, remediation of the stressors is suggested for job incumbents using the job analytic perceptions as guides to program development.

Perceptions of Colorado's Selection Process for Clericals

Bill Maier, State of Colorado Personnel Department

Background

In the spring of 1982, a new selection device for clericals, the "Evaluation of Clerical Training and Experience" replaced the written objective examination give by Colorado up to that time. The previously used written objective test measured standard clerical skills such as spelling, grammar, checking, and computation. The new clerical T&E is a self-evaluation of training and experience across eighty-nine tasks clustered under eleven factors. Applicants self-rate on a 5-point scale which range from "(0) - I have no experience doing this task" to "(4) - I have trained and/or supervised others doing this task." The limited study of validity indicated that the clerical T&E was a valid predictor of typing scores and written test scores.

One of the difficulties of using a self-rating instrument is a tendency of some individuals to inflate their ratings. In order to deal with this problem, an inflation scale was incorporated into the T&E. Non-existent but nice sounding tasks were included and a method of reducing scores based on inflated self-ratings of these tasks was developed (Anderson, Warner, and Spencer, 1984).

This current survey studies the perceptions of the clerical T&E. The survey is intended to investigate the perceived value of the various objectives of clerical selection, the perception of how these objectives were actually being met by the clerical T&E compared to the previously used written, and perceptions on the effectiveness and acceptance of the inflation scale.

Analysis of the Questionnaire

There are four basic parts to the questionnaire. The first part, #1 through #7, investigates the value placed on each of seven objectives of the clerical selection process by the respondents. The second section, #8 through #19, explores the respondents perception of how well the clerical T&E meets the seven objectives compared to the previously used written examination. The third part of the questionnaire, #20 through #23, analyzes perceptions of the inflation of the clerical T&E. Finally, the fourth part of the questionnaire, #24 through #36, defines the characteristics of the raters.

Value of the Objectives

The respondents rated each of the seven objectives in importance on a 5-point scale which ranked from very low to very high. There are three groups of objectives which are significantly different at about the .01 level. The highest group consists of a single member. Validity is significantly more important than quick referral and fairness, which are tied for the second rank. Four objectives from the middle group, quick referral, fairness, objectivity and administrative efficiency, are not significantly different from each other. Finally, the lowest group has two members, Public Relations and Affirmative Action. The most important finding is that validity is perceived as the most important objective of clerical selection.

Comparison of the Clerical T&E with the Written Test

A series of questions were designed to measure perceptions of how the clerical T&E actually performed, compared to the previously used written examination. The summary question allows respondents to rate their overall level of satisfaction with the clerical T&E, compared to the written test. A 5-point scale was used with very satisfied (2) as the top of the scale, neutral (0) at the center of the scale and very dissatisfied (-2) at the bottom of the scale. The average level of satisfaction was significantly above the neutral point for the 109 respondents with a mean of .7. Thus, overall the respondents were more satisfied than dissatisfied with the clerical T&E considering all of the objectives. Each of the seven objectives were evaluated on their accomplishment by the clerial T&E, compared to the written objective examination previously used. Three objectives, administrative efficiency, speed of referral and applicant satisfaction, were perceived by respondents as being significantly better achieved by the clerical T&E than by the written test. The remaining objectives were not perceived as differing between the two examination processes.

Questions 8 through 11 were included in the questionnaire to determine any perceived change in the quality of clericals hired from the clerical T&E. No significant difference in skill level was perceived by the respondents between the clerical T&E and the previously used written test on any of the four specific skills which included willingness to do lerical work, organization skills, typing skills, and clerical abilities, such as spelling, grammar and arithmetic. This is the same lack of significance observed for validity (overall job performance). In summary, the respondents are generally satisfied with the clerical T&E, and feel that it is better in meeting the objective of administrative efficiency, speed of referral and applicant satisfaction.

The Inflation Scale

In developing the clerical T&E, the selection center devised an innovative method ›f controlling the natural tendency of applicants to inflate their training and experience. Respondents placed a very high importance on controlling inflation of application ratings. Eighty-six percent of the respondents rated the importance of controlling inflation as high or very high, while only twenty-seven percent rated it low or very low. The initial telephone survey and unsolicited reports indicated that some clericals knew of the inflation scale and beat the system This contention was supported by questionnaire results, especially for current state clericals now are being promoted.

PAPER SESSION

Psychometric Issues

Chair: Karen Coffey, California State Personnel Board
Discussant: David Friedland, Friedland Psychological Associates, Los Angeles

Ph.D.s Can't Interpret Correlations Either - or Fables I Have Found in the Psychometric Literature

Charles B. Schultz, Washington State Department of Personnel

Psychologists often contend that measures that correlate highly can be used interchangeably. This is due to the fact that we are so used to working with low validity coefficients that correlations accounting for 80% of the variance seem tantamount to perfection. Attributing only 20% of the variance to error of measurement is a happy situation.

Using the same line of thinking, if we find two ways of expressing scores that correlate between .89 and .95, we tend to consider them as interchangeable. However, if the two variables correlate .90, one may have the validity of .10 and the other .52.

Research concerning the use of unit weights vs. regression weights and validity generalization suggest that validity coefficients can be generalized across tests and prediction settings. In many cases, interchanging methods results in a great loss of predictive ability, as in the case where the correlation between the measures is spurious. Also, lack of sensitivity may occur in substituting unit weights for regression weights.

If two equivalent forms of a test correlate .85, we assume that these tests measure the same content but have different errors of measurement. There is another case, the part-whole correlation, in which we measure the same errors but different content.

The correlation itself does not tell us whether we are measuring the same errors or the same content. When a high correlation reflects the same content, the two measures should have similar validity. When a high correlation reflects measuring the same errors, validities can be very different.

Validities of highly correlated tests can differ by .40 or more. The square of the intercorrelation is taken as the proportion of variance in common. Nineteen percent of the variance in Y is free to relate to a third variable, Z, which may correlate zero, with X. When X and Z correlate zero, the correlation of Y and Z can be as high as the square root of .19. Therefore, Y and Z can correlate .44, which may be either positive or negative.

As more and more of the variance of Z become encumbered with the variance of X, the less Z is able to vary independently of X and less of the variance between Y and Z can be independent of X. When X's validity for predicting Z is as great as .50, Y's validity can still exceed it by .33, that is the validity of Y can be as high as .83, but equally important, can be as low as .07. Therefore, our measures may have meaningful differences even though our statistics did not show reliable statistics.

Cultural Bias: Fact or Fiction?

Christina L. Valadez, Washington State Department of Personnel

Washington State has been concerned about resolving questions of test bias for well over a decade. We have recently made an effort to consolidate the results of several of our studies and to compare them to national studies. This work has stimulated interest in considering new ideas for determining the nature and source of cultural bias.

In 1973, a student intern studied selection bias in a multiple-choice test we were using for clerical jobs. Normalized difficulty levels of minorities and Caucasians were compared and showed alphabetization and vocabulary sections of the test to be comparatively easier for Caucasians. Minority candidates did relatively better on the arithmetic section. Other sections showed few differences between groups. This information was used to evaluate job-relatedness of test content when the test was revised.

In 1974, a study was done on "Test Bias in the Caseworker I and II Written Examination Towards Asian-American Applicants." Reading levels and difficulty levels of each item and each section were compared for Asians and Caucasians. The results of this study showed similar answer patterns for both groups.

In 1978 specific multiple-choice items, with difficulty levels that showed a disparate effect on either minorities or Caucasians were selected. These items were compiled into an experimental test and presented without answers. Caucasian and minority state employees who participated in the study were asked to fill in the blank to show how they would handle the situation described. Data on their work performance and general test-taking ability were also collected.

Several staff members have worked on this project when time permitted, with no initial information about the test takers. Now that answer categories responding to the key have been determined, the results are being analyzed. So far, item analyses do show differing response patterns between the minority and Caucasian groups on some items. Several of those items show bias in the same direction as on the original exam. Work performance ratings do not explain these differences. However, initial review of those items shows no obvious reasons why one or the other group would score better on any given item. We are now developing hypotheses about the nature of those items to test out on other exam results.

Test revisions have also yielded some information on cultural bias. In 1983 a staff member begar working on discovering the reasons behind adverse effect on the Employment Security Interviewer I test before beginning to revise it. An item analysis was done for Asians, Hispanics, and Blacks. Difficulty levels for each item were computed and compared. In comparing the difficulty levels, differences between groups on specific items are striking and show reasons for considering different groups separately. For example, an item on interviewing showed a general difficulty level of .84. However, for both Asians and Blacks it was .77; for Hispanics, .95. Work is being carried out now to obtain performance ratings on applicants and compare answers using this criteria. The purpose behind the study is to identify which items have adverse affect, and why, before beginning to revise the test.

Presumably, by identifying and analyzing items that seem to contribute to the differences between minority and majority scores, we can eliminate them from the new test. If we can begin to uncover the causes of adverse effect whether based on cultural bias or other factors,.we will be better able to address the problem in the test development process.

In working toward this goal, we have also been analyzing another test that is being revised. In comparing the mean, standard deviation, and difficulty levels there are again some noticeable differences in answer patterns between the ethnic categories for some of the items. For example, an item asking about the best practice for giving an employee feedback on work performance showed the following results. Sixty percent of the Caucasian sample chose the keyed answer, (a) immediately check with an employee whose work is beginning to decline. The percentage of Asians and Hispanics choosing this response was the same. However, 77 percent of the Black candidates chose this response. Distractor (b), frequent praise for a worker whose work barely meets standards, was chosen by twice as many Hispanics as any of the other groups. This was the least favored response of Caucasians. Distractor (c), withhold comment about work decline until you find out why, was chosen more frequently by Caucasians, but did not attract many responses in the other three groups. Distractor (d), tells employees that no comment means work is satisfactory, was least favored by Blacks and Hispanics.

However, we still don't know what this means. Do these responses represent different group views on how to solve a problem or are they simply individual chance differences? Some studies carried out nationally have attempted to address such questions.

Our data are, of course, not nearly as extensive nor as sophisticated as many of the comprehensive statistical studies that have been carried out nationally. But, based on what we <u>do</u> have, our information compares as follows:

Results of national studies show that tests are equally valid for all ethnic groups in terms of predicting job performance. One instance in which we have validity coefficients for Caucasians and minority groups shows some differences. In this case, the supervisory ratings placed 21 out of 59 Caucasians in the top overall 40 percent. Six out of 11 minorities were in this top group. Yet on the pretest scores, only one of the 11 minorities scored in the top 40 percent; 27 of the 59 Cascasians scored in the top 40 percent. although this differs from the findings of national studies, the small sample size makes it nonconclusive.

The above assumes accuracy and reliability of the performance criterion, as has been noted by Linn and Werts (1971). However, many of our recent validity studies show reasons to question this assumption using either supervisory ratings or an objective performance criterion. Particularly in the case of supervisory ratings, studies have shown reason to doubt this assumption, as factors such as ethnicity of rater and ratee, and similarity in background have been shown to affect perceptions about an individual's capabilities.

National studies show that differences between groups are apparent in different test segments. These differences are borne out as valid when compared to performance differences. Our data also show differences between segments. We have no performance data for direct comparison, but data are being collected to do such a comparison.

According to national studies, variation in test scores due to problems with particular items is disproven. Our item analyses do show differences in response patterns to specific items. However, the differences <u>may</u> be largely accounted for by random sampling errors. They have not yet been shown to be consistently related to anything else.

The conclusion of the national studies is that there is no cultural bias in tests. However, from our perspective, there may or may not be cultural bias. In any case there is definitely an adverse effect, i.e., large differences in test scores. The reasons for those differences have not been fully explained.

Given both national results and our own studies, what can we say we know about bias in testing? We do know that there are differences between groups in overall test scores. We know that group differences are often discernible in sections of a test. We are told that in some cases, at least, the test score differences reflect real on-the-job performance differences.

We don't know, however, how extensive any on-the-job differences are. We don;t know if there are certain problem items that consistently contribute to variation in test scores. And, the big question that we still don't know the answer to is: What accounts for the differences between groups that consistently show up in so many studies?

Earlier attempts at using preventive measures to resolve this problem were based on certain assumptions about the nature of the problem. It was generally assumed in the beginning that situations and therefore test items involving bias could be easily identified by the affected parties. People were asked to identify "culturally-biased items." This usually resulted in identification of items that were difficult for everyone. Protected group members were sought out to particiapte in test development. Yet one of our tests that shows the greatest differences between ethnic groups was developed by an ethnically diverse panel. Language was assumed to cause difficulties. But when few tests were published in a language other than English, candidates tended to receive parallel scores in both languages.

We have eliminated obviously biased situations. We have eliminated questions such as, "What does it mean when you say that you are looking at the world through rose-colored glasses?" We have de-jargonized and de-fogged. We have made tests clearer and more job related. Yet none of these measures has been totally successful in achieving the desired result. Perhaps, then, it is time to begin to reexamine our assumptions about the nature of "cultural bias."

Development of a Mathematical Model to Estimate the Equivalence of Repeated Administrations of a Measurement

William E. Donnoe, California State Personnel Board

At the California State Personnel Board job applicants are evaluated by means of written tests and/or interviews. Rather than examining people for vacancies that exist, the State Personnel Board examines people to establish lists of eligibles from which future vacancies can be filled. When examining large numbers of candidates for placement on a single list of eligibles it is often necessary to use more than one panel of raters to interview candidates. As with any measurement, questions regarding the reliability of the measurement exist.

In interviewing candidates for placement on employment lists the State Personnel Board uses a relatively standardized process called Qualifications Appraisal Panel interviews, or QAP's. These QAP examinations use a predetermined set of evaluative criteria specific to the job classification being examined. These criteria outline the factors on which candidates are to be competitively rated and accordingly ranked for placement on a single list. Panels typically consist of one representative of the State Personnel Board and one or two panel members from the State departments that use the job classification being examined. Typical interview schedules call for interviews to be scheduled 20 to 30 minutes apart, with one panel working straight through for seven to ten days maximum. Where the size of the candidate group is too large for one panel to realistically evaluate all candidates, two or more panels would be used and the results of the entire process combined to form one list of eligibles. Some of these panels interview candidates at one location only and other panels may travel to a number of location throughout the State to meet with candidates (which may be a factor in scoring patterns).

Candidates are scored independently by each panel member with the candidate's final score being the arithmetic mean of the raters' scores. Panel members are encouraged to discuss their perceptions and scoring of each candidate prior to assigning final scores. Panel members rarely deviate in their final scores for any one candidate.

Problems arise when it comes to the attention of the State Personnel Board that different panels are coming up with different score patterns (i.e., one panel giving predominantly high scores and one giving predominantly low scores). Where differences between panels occur it can be attributable to two sources: First, such variance could be due to the quality of candidates seen by the different panels; or, such variance could be attributable to the way in which the different panels used the rating criteria (one panel being "easy" on candidates, one panel being "hard" on candidates).

By assuming that all candidates are assigned to panels on a quasi-random basis, it can be hypothesized that the spectrum of abilities to be evaluated by each panel adhere to a model of randomly distributed abilities. As such, panels should be confronted with a heterogeneously distributed array of abilities among panels.

At the time at which this study was undertaken, the State Personnel Board evaluated such conditions by looking for significant differences. First among panels: Did significant differences in scores among panels appear? If so: Could this be attributable to significant differences in the candidate groups seen by these panels? Where differences occured, and these could not be attributed to differences in candidate groups, it was assumed that variance among groups was due to differences in rating approaches. The reliability of the administration of the test was being questioned with no estimate of such reliability. Traditional models of reliability were not entirely appropriate for the unique set of circumstances produced by multiple panel interviews.

I will present a model by which the consistency of administration of this type of interview examination can be estimated. By viewing the variance among the interview panels (error variance) and the variance within the interview panels (systematic variance), the model developed estimates the extent to which the observed variance is free of error associated with interview panel assignment. This model of test administration reliability (rpp) provides a coefficient of equivalence of administration of a measurement, which can be best described as an alternate-forms model of reliability.

Three recently administered SPB examinations were used to determine the rpp model. The observed rpp values were all indicative of consistenly administered interview examinations (rpp values exceeded .94 for all three examinations). This estimate of inter-panel consistency was extended to an estimate of standard error for each of the three examinations. The observed standard error statistics were interpreted as an indication of the extent to which panel assignment influenced the observed (true) interview scores.

Recommendations were made to the SPB to adopt this model to estimate examination administration consistency. Future applications of this model were forseen for any condition where multiple administrations of a measurement would be combined.

SYMPOSIUM

How to Justify Ranking When Using Content Validity

Chair: Bruce W. Davey, Connecticut Department of Administrative Services

Ranking Candidates Based On Content Valid Tests

Nancy E. Abrams, Personnel Management and Measurement

Ranking of candidates on content validated procedures has become one of the major areas of challenge in court cases today.

Charley Sproule has prepared an article based on his workshop last year for the upcoming IPMA journal issue on selection. The article covers the major parts of the Guidelines and Standards and court cases related to this issue.

> He states that the Guidelines require evidence of a relationship between what is measured by the selection procedure and differences in levels of job performance and also that the closer the selection procedure approximates important work behaviors the easier it is to make that inference.
>
> The Standards encourage the use of confidence intervals in score reporting.

In my opinion, these are two major issues in justifying rank for content valid procedures.

On the first issue, many people have used job analysis scales - i.e., does more of this make someone a better worker or Job Element - does this differentiate superior from barely acceptable workers? These scales may be adequate but many have gone beyond them to the types of methods Bruce Davey and Bill Holland have discussed.

At last year's IPMAAC conference, I have discussed work done in conjunction with determining the passing point for police and firefighter physical agility tests in New Jersey. We asked supervisors to observe candidates taking the test and for each one, select a critical incident which would be their prediction of each person's job performance (all of the critical incidents were physical in nature). In determining the passing point we were most concerned with the test scores which differentiated those rated with acceptable incidents from those rated with unacceptable incidents. In addition, we discovered that those scoring higher on the test were rated with the higher valued incidents, providing justification for test ranking. This type of evidence (as is the evidence Bruce and Bill discussed) is clearer to present in justification of test ranking since a specific prediction of job performance can be made from a particular test score.

In relation to the second issue: test precision, many courts have begun to look at this issue since the Guardian IV decision (Guardian Association of New York City Police Department, Inc. v. Civil Service Commission, et al., 630 F. 2d 79 (2d Cir. 1980). In that case, the court saw the scores bunched at a narrow range of score levels and felt that in relation to the standard error of measurement precise distinction could not be made which resulted in adverse impact. The standard error of measurement, the distribution of scores and the degree of adverse impact created by trying to make precise score differentiations have led courts to dismiss the ranking use of tests. These are important analyses to do on your tests.

In summary, a test user who is basing a test on content validation to rank order candidates must be prepared to show the relation of test performance to job performance and also show that the scores are not being interpreted more precisely than is warranted by the standard error of measurement of the test.

Ranking Candidates

Bruce W. Davey, Connecticut Department of Administrative Services

Undeniably, there are situations which require a pass-fail test versus one designed to rank candidates. For example, there is a specific situation presented in the Uniform Guidelines supplementary section, entitled "Questions and Answers". In the "Questions and Answers" section of the Guidelines, a situation is posed in which an incumbent in a particular job is required to lift 50-pound boxes from point A to point B. The Guidelines make the point that you cannot claim content validity for a procedure which requires candidates to carry weights greater than 50 pounds because the job doesn't require anything greater than 50 pounds, and you might be excluding some great 50-pound toters at the expense of those who are great at 100 pounds. In other words, you might be bypassing people who could do a fine job hauling 50-pound cartons on the basis that they didn't do as well as people who can haul heavier cartons than will ever be required on the job. This is a classic case of an artificial barrier.

A similar example involved secretarial skills -- in particular, typing speed and stenography provide good examples. If a ranked list of stenographers places people at the top of the list who are skilled at taking dictation at 120 words a minute or more, that list is not necessarily appropriate to select a steno for a boss who has never dictated faster than 100 words a minute. Should you rank typists according to their speed if they're going to be in an environment which rarely demands raw speeds greater than, say, 50 words a minute, but always demands accuracy?

Along similar lines, you would have a hard time defending the content validity of a test which measured addition skills ranging from simple to highly complex if the job itself required only that incumbents be able to add single-digit numbers.

I think that for the examples I've just cited, an argument can be made that ranking candidates is inappropriate. However, you may have noted that all of the examples given involved skills with fairly discrete performance levels. They did not involve knowledge. Scores on a test of knowledge cannot be interpreted in the same way as scores on a test of skill.

The skills tests that I've described have all been set up so that scores on the skills test reflect different skill levels. However, a properly-developed content-valid test doesn't concern itself with multiple levels. A truly content-valid test concerns itself only with the level of knowledge or skill which is relevant to performing the tasks of the job -- not with performing at levels which are not required to do the job. In other words, the test doesn't contain items which test knowledge above the level of the job, or scoring methods which give credit for skill levels beyond that required for the job.

I need to differentiate between knowledge and skills in order to make my point a little better. Skills tests generally lead to a score which denotes a specific level of skill, and it's usually something which can be objectively quantified. For example, typing speed is measurable skill; you can type at 50 words a minute, or 60 or 80 or 100. Steno speed falls in the same category. The example involving package-lifting involved discrete amounts such as 50 pounds or 100 pounds, and could have involved points between, such as 80 or 90 pounds. The point is that all these measurements involve identifiable levels, and if you've achieved that level you're better or faster or stronger than people below you.

But, in my opinion, a properly content-validated test involves only one level -- the level of knowledge or skill required to do the job. It does not -- or should not -- test levels beyond that required to do the job. A properly validated typing test will not rank candidates above the level required to do the job. A properly validated box-lifting test will only deal with job-related materials, and therefore will only deal with carrying boxes with weights similar to those encountered on the job. In other words, the test has to be at the *appropriate difficulty level* for the job, as required by the *Uniform Guidelines*. A true content-valid test wouldn't even *ask* candidates to carry a one hundred pound box if the maximum job requirement was 50 pounds. The issue isn't so much an issue of ranking as it is one of *relevance*. Thus the *Guidelines* have used a self-serving example.

That last point is important because it has a bearing on the argument I want to advance regarding ranking candidates on tests of knowledge as opposed to measurable skills. My opinion is that if you have properly documented the content validity of each question on a job knowledge test, then the test you have developed will be appropriate to use a ranking device. Maybe it won't be appropriate from the perspective of the *Uniform Guidelines*, but it *will* be psychometrically appropriate.

That is my key point. I seriously question whether there's aneed to separately establish ranking validity when you have a test composed entirely of items which a panel of experts has rated as covering things that incumbents *should* know. Such a test isn't asking anyone to lift 100 pounds when the job requires 50 -- and it isn't asking anyone to call up from memory some archaic or little-used bit of knowledge or some fugitive from the game of Trivial Pursuit -- it is asking only those items that most incumbents *should* know. And doesn't it seem self-evident that when you're talking about a test of essential job knowledge, that the incumbent who knows more of the basics or the essentials can justifiably be ranked higher than someone who knows less?

Well, that's basically how I feel about demonstrating ranking validity for content valid tests. But nevertheless, the Uniform Guidelines tell me that I must put these feelings aside and validate for ranking. So last year some time, I began the search for methods to document ranking validity. Along the way, I found out two things. I found out that there didn't appear to be too many formal systems for meeting this requirement, and I found that most of them were tied to the analysis methodology rather than to the review of test material. But, given that I had to meet the ranking requirement, I was much more interested in a system that aimed at documenting the capabilities of the test itself to rank-order candidates appropriately. And the best system I found for this purpose was the one developed by Bill Howeth of McCann Associates.

Since Bill has already presented his system, I don't have to go into much detail on it. I'll just go ahead and show you an adaptation of it which I developed in conjunction with a test development project:

1. The Connecticut system was designed to be easy for the raters to follow and easy for the analyst to compile -- by multiplying ratings by 10, one has the raters' estimate of what percentage of the best, middle and worst

performers are expected to do on this item.

2. The end product of the system is a collection of questions which in combination can differentiate between candidates' levels of job-related knowledge. Using the ratings collected, it can be mathematically shown whether the experts used as raters feel that the top performers of the job would do better on the test than the middle or bottom performers. In fact, the data can be used to make an estimate of the criterion-related validity of the exam.

An Analysis of Legal Issues Affecting Content Validation

Patrick T. Maher, Personnel and Organization Development

Content validation has been recognized as an acceptable form of validation in Title VII cases since the inception of the original EEOC Guidelines. While recognizing content validation, the EEOC actually favored criterion validation. Some of the earliest court decisions involving content validation cases seemed to favor criterion validation while giving content validation only grudging recognition. In most of these cases, however, the validation studies and resulting assessment procedures were "so artlessly developed" that the courts readily rejected them. In doing so, however, there was a tendency to criticize content validation generally, and in some cases, to include comments that tended to either reject content validation or to effectively render it inappropriate for other than the most simple physical jobs (i.e., the classic typing example).

A review of federal appellate court decisions on the issue of content validation in Title VII cases has revealed two trends. One is that content validation has not been substantially litigated at the appellate level. The second trend is that in cases in which content validation has been a significant issue seem to indicate that the courts are willing to accept content validation as an acceptable form of validation in Title VII cases, even when the jobs that are involved are not necessarily physical and even if they do not necessarily result in tangible products.

Where content validation has been criticized, especially recently, the courts have not rejected content validation per se, but instead have found fault with the issues that are specific to the facts of the case at bar.

Legal Status. The courts have repeatedly upheld content validation as an acceptable form of validation in Title VII cases, even in those cases where the courts seemed to strongly favor criterion validation or seemed to set such strict standards for content validation that they seemed impossible to meet, at least in theory, content validation is legally recognized in Title VII cases.

Job Analysis. The courts mandate a thorough job analysis as a prerequisite to content validation as well as to test construction based on content validation. While the courts prefer that each KSA be linked to a specific work behavior, they will not necessarily reject the job analysis if this step is missing. In such circumstances, however, the courts will look at KSAs and determine if they can make a logical linkage between the two.

Job Relatedness. Some early decisions described "job relatedness" as activities or factors that were imperative to the operation of the agency. The courts have backed away from this strict standard and are now defining job relatedness in terms of using "professionally accepted methods" to show that tests measure traits that are "predictive of or significantly correlated with important elements of work behavior that are relevant to the job for which the candidate is being evaluated."

Cutoff Scores. The courts recognize that in any decision-making process that relies on a standardized test, there is some risk of error, no matter what the cut-off score is. Consequently, content is some independent basis for choosing the cutoff. The courts are willing to recognize such factors as candidate population versus number of openings, civil service rules establishing minimum scores, or even analyzing test results to locate a logical breaking point in the distribution of scores.

Ranking. Some early appellate court decisions, and some more recent decisions at the trial level have tended to hold that ranking of candidates is not permissible using content validation if a Title VII issue arises. Recently, however, the appellate courts are changing their view and are willing to accept the idea of ranking under content validation.

In Summary, Content validation is alive and well and seems to have a growing acceptance. While the courts are not going to require strict adherence to every single professional and legal standard governing content validation, they are going to require substantial compliance with them. The more consistent any specific procedure is with professional and legal standards, and the more the procedure involves simulation testing, the greater the likelihood that the courts will accept content validation.

Following are the major decisions at the federal level, with emphasis on the appellate level, that deal with content validation. Some cases do not specifically address content validation, but have been included because the courts have specifically referenced them in discussing content validation in other cases. Significant content validation cases are marked with an asterik.

Albermarle Paper Co. v. Moody, 422 U.S. 405, 95 S.Ct. 2326, 45 L.Ed.2d 280 (1974)

Association Against Discrimination in Employment v. City of Bridgeport, 594 F.2d 306, (2nd Cir., 1979) on remand 479 F.Supp. 101, aff in part, vac in part 647 F.2d 256, cert den Bridgeport Firefighters for Merit Employment v. Association Against Discrimination in Employment, Inc., 102 S.Ct. 1611, 455 U.S. 988, 71 L.Ed.2d 847

Bridgeport Guardians, Inc. v. Bridgeport Civil Service Commission, 482 F.2d 1333 (2nd Cir., 1973)

Bridgeport Guardians Inc.v. Bridgeport Police Department, 431 F.Supp. 931 (D.Conn., 1977)

Chance v. Board of Examiners, 458 F.2d 1167 (2d Cir., 1972) affirming 330 F. Supp. 203 (S.D.N.Y., 1971).

*Contreras v. City of Los Angeles, 636 F.2d 1267 (9th Cir., 1981) cert den 102 S.Ct. 1719, 455 U.S. 1021, 72 L.Ed.2d 140

Craig v. County of Los Angeles, 626 F.2d at 659 (9th Cir., 1980)

Crockett v, Green, 388 F.Supp. 912 (E.D. Wis. 1975), aff'd, 534 F.2d 715 (7th Cir., 1976) (Disapproved in Contreras).

DeLaurier v. San Diego Unified School District, 588 F.2d 674

Espinoza v. Farah Manufacturing Co., 414 U.S. 86, 94 S.Ct. 334, 38 L.Ed.2d 287 (1973)

*Firefighters Institute for Racial Equality v. City of St Louis, 616 F.2d 350 (8th Cir., 1980), cert den City of St Louis, Mo v U.S., 101 S.Ct. 3079, 452 U.S. 938, 69 L.Ed.2d 951

Firefighters Institute for Racial Equality v. City of St Louis, 588 F.2d 235, cert den Banta v Firefighters Institute for Racial Equality, 99 S.Ct. 3096, 443 U.S. 904, 61 L.Ed.2d 350, cert den City of St Louis, Mo v U S, 101 S.Ct. 3079, 452 U.S. 938, 69 L.Ed. 2d 951

Firefighters Institute for Racial Equality v. City of St Louis, 549 F.2d 506 (8th Cir., 1976), cert den 434 U.S. 819, 98 S.Ct. 60, 54 L.Ed.2d 76 (1977)

Guardians Association of New York v. Civil Service Commission, 431 F.Supp. 526 (1977) 526, vac and remanded 562 F.2d 38, on remand 466 F.Supp. 1273

*Guardians Ass'n of New York City v. Civil Service, 630 F.2d 79 (8th Cir., 1980) cert den 101 S.Ct. 3083, 452 U.S. 940, 69 L.Ed.2d 954, on remand 527 F.Supp. 751

Jackson v. Nassau County Civil Service Commission, (E.D.N.Y., 1976)

Kirkland v. Department of Correctional Services, 374 F.Supp. (S.D.N.Y. 1974), aff'd in part and rev'd in part, 520 F.2d 420 (2d Cir. 1975), cert den 429 U.S. 823, 97 S.Ct. 73, 50 L.Ed.2d 85 (1976)

McDonnell-Douglas Corp. v. Green, 411 U.S. 792, 93 S.Ct. 1820 (1973)

Richardson v. McFadden, 540 F.2d 744 (4th Cir., 1976)

Tyler v. Vickery, 517 F.2d 1089 (5th Cir., 1975), cert den 426 U.S. 940, 96 S.Ct. 2660, 49 L.Ed.2d 393 (1976)

United States v. Georgia Power Co., 474 F.2d 906 (5th Cir., 1973)

United States v. North Carolina, 400 F.Supp. 343 (E.D.N.C. 1975), vacated on other grounds, 425 F.Supp. 789 (E.D.N.C. 1977).

United States v. State of South Carolina, 445 F.Supp. 1094 (1977)

Washington v. Davis 426 U.S. 248, 96 S.Ct. 2040.

The following sources were obtained from court decisions that made specific reference to them when discussing content validation issues. Because of this, they tend to have legal support and should be reviewed and considered in using content validation.

APA Standards, at 48-55.
Catell, Validity and Reliability: A Proposed More Basic Set of Concepts, 55 J.Ed.Psych. 1 (1964).

Ebel, Comments on Some Problems of Employment Testing, 30 Personnel Psych. 55 (1977).

Science Research Associates, Validation: Procedures and Results (1972) (use of criterion "tails" identifying best and worst candidates more justifiable than continuous rating).

Tenopyr, Content-Construction Confusion, 30 Personnel Psych. 47 (1977).

PERSONNEL TESTING COUNCIL INVITED SPEAKER

"Performance Measures: Forms or Samples"

Dr. Sheldon Zedeck, University of California at Berkeley

Overview. I chose the title. and posed the question -- Performance Measures: Forms or Samples? -- with an answer in mind. That is, samples are the better way to measure performance and what better way to sample behavior in a relatively standardized fashion than via assessment centers. We have always believed that the best predictor of future performance is past performance. The data are quite impressive showing that assessment center exercises are good predictors of future performance. My major point will be that if they can measure or predict future performance, they should be able to accurately measure current performance. What I'd like to do for this presentation is justify this position on the basis of logical and empirical grounds -- grounds that are not always overlapping.

The essence of my talk will be on behaviors. Performance appraisal is concerned with the measurement of one's competency. There are several purposes for which we conduct performance appraisals. Since we want to use the performance appraisal data for different purposes, we will make inferences from one's measured competency where these inferences may vary depending on the purpose. And since we are into inferences of results, we are basically concerned with the validity of our performance appraisals. I will argue that the best way to establish the validity of a performance appraisal is to develop samples of behavior and, consequently, have a content valid performance appraisal system.

What's Wrong with Performance Appraisal Forms? The basic question is: What's better -- forms or samples? The reason that I opt for an emphasis on samples is due to the fact that there are many problems with the forms that we currently use. A perusal of the literature shows (1) that all sorts of forms are susceptible to response biases; (2) that there is unwillingness to complete appraisal forms; (3) that there are equivocal reliabilities across different raters and levels of raters; (4) that the forms are limited as information gathering devices and thus limited in their value as an aid in information processing; and (5) etc. It is not that I am completely against forms, but rather, that we have spent too much time, effort, research, and money on looking for the form. Though the last few years have seen increasing emphasis on the performance appraisal process, recent books still devote about 25% of their contents to methods and methods-related issues (e.g., See Bernardin & Beatty, 1984). I realize that any evaluation process that requires documentation also requires a form. My point, however, is that we have placed too great an emphasis on the form or the method.

I see a distinction between what I refer to as static forms as opposed to dynamic methods, procedures, or processes. BARS, graphic rating scales, forced choice, BOS, checklists, and the like are examples of static forms, the intent of which is to gather data in order to determine a numerical value for a ratee that summarizes his/her performance. All of the forms are usually the same for all of the ratees. The anchors, behaviors, statements, etc., are used to serve as benchmarks for the rater; none of the items or statements are changed or adapted as one goes from one ratee to another or even across time. Dynamic processes, however, are systems that allow for the accumulation of data, in the form of behavioral examples, and which are generally obtained such that they are directly pertinent to the ratee; i.e., the data are unique to the ratee's performance over the period of time for which one is doing the evaluation. Numbers may be attached to these data, but the concern is on the descriptive data and not on the numerical values attached to them,

I believe that different appraisal purposes require different methods. Yet, I don't think many organizations use different forms for different purposes. Nevertheless, I'd like to review some of the purposes. A list that I have used (See Jacobs, Kafry, & Zedeck, 1980) and that has been adapted by Bernardin and Beatty (1984) is one that contains the following purposes:

(1) Feedback/Employee Development -- use of appraisals to provide concrete and specific feedback to employees.
(2) Promotion, Merit Pay, Placement, and Disciplinary Action Decisions--These were clustered together by Bernardin and Beatty (1984); Jacobs et al. (1980) had them as separate purposes.
(3) Selection Research--Appraisal data used as criteria in test/validation studies.
(4) Training/Supervision--Appraisal data to develop training curricula and to determine training needs.
(5) Organizational Diagnosis and Development--use of appraisal data to detect organization-wide problems and manpower deficiencies and to set performance goals at the organizational or unit level.

From an operational perspective, can a single form be designed to meet the needs of all of the above "utilization" functions? The answer is: not very likely. Particular forms are needed for particular purposes. There needs to be a match between form and purpose, yet one form is usually the sole evaluation form in the organization from which it was taken.

To continue, if different purposes require different forms, then perhaps different purposes require different strategies for evaluating personnel. A while back, Wayne Cascio and I examined such issues (Zedeck & Cascio, 1982). In particular, we created 33 hypothetical profiles of a supermarket cashier and asked subjects to evaluate the performance of each of the hypothetical checkers on 7 point scales. There were three groups of raters, each group evaluating for one of the following three purposes: (1) recommending development; (2) awarding a merit raise; (3) retaining a probationary employee. Thus, each of the three purpose group subjects examined and evaluated the same stimuli -- the 33 hypothetical descriptions of checker performance. Analyses revealed the following:

(1) If we concentrate on the decisions that were made, we find that there were significant differences among the three purpose groups. 19% of the decision variance (variance in ratings across groups) was explained by the purpose manipulation. Thus, there is relatively strong support for the conclusion that evaluations differ as a function of purpose.

As an aside, I submit that these data put a damper on much of the research undertaken in the area of performance appraisal. My review of the literature shows that most studies ask the respondent to evaluate the ratee's performance without any reference to performance for any particular purpose. Our data show that you will evaluate the same ratee differently depending on whether that ratee is being evaluated for a merit increase or for retention. What purpose was being implicitly considered by the rater in much of our research studies?

(2) If we examine the specific information used to influence the ratings, not only were the decisions different for the three groups, but policy capturing analyses indicated different strategies of evaluating within and between purpose groups. With regard to between purpose group differences, those evaluating for merit pay keyed primarily on performance in the domains of "skill in human relations" and "organizational ability." Those

results suggest that rater strategies and information processing capability vary with the purpose of the rating. Identical performance dimensions (domains) and performance profiles are weighed, combined, and integrated differently depending on whether the purpose of the rating is for a merit raise or for development or retention. The point is that raters differ in how they evaluate -- they come to different conclusions/decisions not only as a function of purpose, but also because of individual differences.

In sum of the above, I submit that the data argue against using the same form for all purposes and I will further submit that an emphasis on forms is not that suitable for particular purposes.

I mentioned individual differences among raters. Let me pursue this issue. Taft (1955) showed that there is variability in the ability to judge people, particularly when the judges are clinicians. Extending this logic to performance judgements, Zedeck and Kafry (1977) studied whether certain components of the evaluation process are influenced by individual differences in observation skills. In particular, they studied the relationship between certain individual difference variables and (1) raters' ability to accurately identify performance levels for particular behaviors, and (2) the degree to which judges (raters) could accurately allocate behaviors to performance dimensions. Both of these goals are consistent with today's research trends to study the cognitive aspect of the performance process (see Cooper, 1981; Feldman, 1981).

The results showed that those who scored higher on measures of intellectual efficiency and verbal reasoning, and who were task-oriented on leadership scales were better at evaluating and categorizing behaviors. Furthermore, those who were more "perceptive" as measured by a "social insight" test were better able to distinguish dimensions. In sum, these results suggest that there are individual differences in the ability to perform certain functions in the performance appraisal process. If raters vary in their ability to perform certain appraisal functions, then perhaps a consensual process would be the optimal solution.

My concern for individual differences and cognitive processing in the appraisal task has continued. This past year, Sharon MacLane and I conducted a laboratory study concerned with identifying those who view or organize behavior into "fine" or "gross" components or units. For example, I might turn, walk over, flip the switch to turn on the lights, and walk back to this spot and you might see each of these actions as separate, meaningful actions. Or you might see them as just one action, such as "turning on the lights in this room." The former view is that of a "fine" analyst whereas the latter perception is that of a "gross" analyst. This may not seem like a real or important distinction, but there are data in social cognitive psychology that show that differences in levels of analysis -- fine vs. gross units of analysis -- relate to amount of information gained from observation (Newtson & Rinder, 1979). The "fine" analysts are those who see more break points in behavior and consequently summarize, integrate, and store information differently than do the "gross" analysts.

In our study, we had subjects observe a 20 minute lecture in a classroom (video tape) of either a good lecturer or poor lecturer. Subjects were instructed to push a button attached to an event recorded each time they observed behavior that was meaningful to them for evaluation purposes. It was our supposition that more "button pushes" represents "fine" analysts; fewer "button pushes" represents "gross analysts." The dependent variables in our study were tests of recall and recognition as well as rating accuracy.

I will only describe some of the "gross" results. In general, we found that the number of button pusheswas significantly correlated with the score on a true-false test of what was observed (r=.22;n=60;p<.05). Those who had more "button pushes" scored higher on the true-false test concerned with what was seen on the tapes. Furthermore, on each of seven rating scales, those who were more "gross" tended to provide more lenient ratings; i.e., those with less "button pushes" evaluated the performance more highly. There were no differences, however, between the "gross" and "fine" raters in their confidence of rating or between the good or poor performance (lectures) tapes. These data further support the notion of individual differences, particularly with regard to the degree to which judges observe, store, and categorize behavior.

Can we train people to be better observers of behavior; i.e., to be more accurate. My own research indicates that training is not too effective (Zedeck & Cascio, 1982). A review by Spool (1978) of 25 years of research, however, is more positive, but nevertheless, still equivocal.

The above research findings are disturbing to me since we continue to look for *the* method or form. Concentration on developing *the* method will not result in the situation where we will automatically obtain accurate ratings. There are simply too many factors that influence ratings, and the one domain of factors that I have emphasized today is that of individual differences among raters. It is my contention that we would be better served if we de-emphasized forms and concentrated on processes that allowed us to sample behaviors and to evaluate these behaviors based on a consensual process.

Samples of Performance

Assessment for Selection Purposes. The literature often serves as the basis or catalyst for our proposals and the present case in no exception. My "classic" reference is the article by Wernimont and Campbell, in 1968, entitled "Signs, Samples, and Criteria." One of their major points was that we should focus on meaningful samples of behavior rather than on signs of predispositions (which I take to be the numerical values provided on performance appraisals). As I previously mentioned, the axiom is: "The best predictor of future performance is past performance." They proposed a behavioral consistency model that requires concentration on dimensions of actual job behaviors. Reliance on the consistency notion forces a consideration of which job behaviors are recurring contributors to effective performance and which are not.

The research that I am going to describe deals with samples of behavior that are used in selection and in development. Most of the usual correlational data that often are reported in selection studies will also be presented here; correlations pertaining to the developmental situation are not currently available since the study is still on-going, and as I will state later, are probably irrelevant to the purpose. In essence, I will be describing how performance measures that are based on samples of behavior can be used to select current employees who are to be promoted as well as to identify developmental needs of current employees.

The position of interest is that of Account Executive in a large financial institution. The purpose of the selection project was to take current employees in the bank and "upgrade or transfer" them--a form of promotion. Basically, an account executive is responsible for identifying, proposing, and explaining loan and savings products that are suitable for the bank's customers. In addition to the basic task of selling, account executives need to prospect new clients, research market and client information, and provide customer service. These domains, as well as the information to be used to develop all aspects of the behavioral system, were identified via a reasonably thorough job analysis.

The three major samples of behavior gathered in the selection project are: (1) the supplemental application form; (2) an interview/role play; (3) assessment center exercises.

1. The Supplemental Application Form (SAF). The SAF is an application blank that is consistent with the models proposed by Hough (1984)--the Accomplishment Record--and by Schmidt et al. (1979)--the behavioral consistency method of unassembled testing.

The SAF is a form that requires candidates to describe previous experiences, achievements, and accomplishments that deal with job relevant dimensions. For the account executive, we determine that there were 6 relevant dimensions: (1) Acquiring Information; (2) Decision Making and Judgement; (3) Organization and Planning; (4) Sensitivity; (5) Behavior Flexibility/Goal Oriented; and (6) Persuasiveness/Professional Impact. Separately, for each of these areas, the candidates were required to indicate when and how they demonstrated their skills in the dimensions. Examples or incidents could come from financial institution experiences or any other experience--volunteer work; school activities; etc. It was suggested that, at most, two examples should be provided for each dimension.

A group of personnel and line people then were trained to score the SAF. Criteria or benchmarks were established by which to evaluate each SAF. Each SAF, with the name of the ratee removed, was subsequently scored independently by two evaluators. Analysis of the scores of the SAF, in part, were used to de-select about 20% of the applicants.

2. Interview/Role Play. Successful candidates on the SAF moved on to a 30 minute interview in which the candidates' experiences were elaborated upon. In addition, there was a 15 minute role play that was concerned with

a simple sell--that of opening an account in the bank. The interview aspect resulted in evaluation on the same dimensions as those evaluated in the SAF; the role play aspect resulted in the evaluation on dimensions such as oral communication, oral defense, attention to detail, persistence, and the like --these dimensions were identified in the job analysis.

Overall success on this component--the interview and role play--resulted in going on to the assessment center simulation.

3. Assessment Center Exercises. Candidates (n=62) were placed into a one day center that contained five simulations: (1) a 2 hour in-basket followed by a 1 hour interview; (2) a loan decline situation where the candidate needs to inform a client (role player) that he/she was denied a certain type of loan, but may be eligible for another type; (3) an oral presentation on a product to a group of potential clients as well as ones who might be able to refer clients; e.g., financial planners; (4) a prospecting exercise--a two part exercise (15 minutes in the morning and 30 minutes in the afternoon) in which the candidate needed to identify a client's (role player) needs and then return in the afternoon having qualified the candidate and proceeds to sell him/her products; (5) a customer walk-in, or "cold sell", in which the customer (role player) has shopped around before seeing the candidate.

Prior to attending the center, candidates (n=62) received a pre-assessment center package that described the fictitious financial institution and its products. During the course of the exercises, candidates were being observed for behaviors that are indicative of the 14 following dimensions:

ASSESSMENT CENTER SKILL DIMENSIONS

1. ORAL COMMUNICATIONS	Ability to verbally convey thoughts and ideas in a clear, unambiguous, and effective manner.
2. ORAL DEFENSE	Ability to verbally explain conclusions and logic underlying their choice in an effective manner.when challenged.
3. WILLINGNESS TO MAKE DECISIONS	Ability to make decisions when needed without reluctance.
4. QUALITY OF DECISIONS	Ability to make high quality decisions based on clear-cut, logical rationale.
5. RESISTANCE TO PREMATURE JUDGMENT	Ability to resist coming to conclusions before collecting and evaluating pertinent information.
6. ACQUIRING INFORMATION (ORAL)	Ability to collect relevant information by questioning and discussion.

7. ATTENTION TO DETAIL

Ability to handle all the details encountered in performing a task; ability to collect the important information relevant to a customer from written sources.

8. SENSITIVITY

Ability to detect and interpret subtle uses in the behavior of others concerning their reaction to a situation and to interpret social cues from others concerning the appropriateness of one's own behavior.

9. ORGANIZING AND PLANNING

Ability to systematically arrange work and establish priorities to effectively accomplish work.

10. BEHAVIOR FLEXIBILITY

Ability to modify behavior, when motivated, to reach a goal.

11. PROFESSIONAL IMPACT

Ability to establish credibility as an expert in solving financial problems.

12. PERSUASIVENESS

Ability to change the thinking and behavior of customers using pertinent data and applying logic without generating resentment.

13. PERSISTENCE

Ability to persevere in attempts to persuade customers to one's own point of view using a variety of supporting comments.

14. RESULTS-ORIENTED

Ability to reach a goal/objective in spite of distractions, interruptions, and conflicting work priorities.

Before I present the results, I would like to note that the three aspects of the process--(1) the SAF; (2) interview/role play; and (3) the assessment center exercises--vary in the degree to which samples of behavior are presented and emphasized. Samples are not necessarily complete job simulations or job replicas; rather they represent behavior that is relevant. In our situation, we had simulations that represented a high degree of "fidelity" to the job (role play interview and assessment center exercises), but also permitted self-reported descriptions of behaviors that were relevant to the job requirements, though not necessarily from a duplicate job situation.

Results. Each of the candidates was discussed in sessions that lasted about 1.5 hours. The purpose of the discussion was to review performance in the center and to evaluate the candidates on each of the 14 dimensions. Each of the dimensions was defined in some detail; in addition, there was an anchored scale for each dimension. The essence of the discussion was on behaviors that were or were not emitted. The discussants were trained assessors who came from managerial positions within the functional line in which sales operated. After consensus was reached on whether the candidate should be placed into a training course.

About two months after being trained, candidates (n=24) were called back into the central office and administered a product knowledge test, a product judgment test, and participated in two role plays that simulated typical and frequent encounters for the account executive. The latter use of role plays in this stage is consistent with the suggestion made by Thorton and Byham (1982) that an assessment center might be an appropriate technique to evaluate training. We used the results, here, in part to evaluate training and as a criterion against which to evaluate the selection data.

Some of the results are as follows:

(1) The overall evaluation on the SAF correlated .26(n=98;p<.01) with an average of the ratings made in the interview on the same dimensions as those assessed on the SAF.

(2) The overall evaluation on the SAF correlated .34(n=98;p<.01) with the overall rating made for the interview session.

(3) The overall evaluation on the SAF correlated .46(n=98;p<.01) with the assessment center consenses overall rating.

In sum of the SAF results, we see that it is a good predictor of assessment center performance. The fact that it is less costly and involves less time and effort on the part of the organization suggests that it might become a valuable and permanent part of the selection process.

(4) The assessment center consensus overall rating correlated better with ratings on the same types of dimensions as measured by the SAF(r=.46) than with the interview ratings on the same type of dimensions(r=.23.). Since the interview is more susceptible to biases or personal idiosyncracies of the interviewer, or to an interviewer-interviewee match, the good result for the SAF is encouraging, and consequently, could replace the interview in the future.

(5) The overall ratings of the job simulation role play performed by the candidates after being on the job for two months correlated with the selection interview dimension ratings (r=.36) and the overall interview ratings (r=.55). It also correlated with the overall performance rating, but in a negative fashion (r=-.37).

In general, the data and results show linkages between the assessment center, interview role play, and SAF ratings--all of which rely on behaviors either emitted in a simulation or assessed based on one's description. In contrast, the typical rating scale does not yield significant correlations except for the anomalous correlation of-.37 between the job simulation performance and overall job performance. The latter measurement was obtained from field supervisors. This correlation indicates that those judged to be better in overall performance were judged to be poorer in the job simulation. The result can be explained, in part, by looking at some of the descriptive statistics for the overall rating: mean=4.39; SD=.94; with a 5-point rating scale, there was a frequency distribution of one"1", one "3", eight "4's", and thirteen "5's"--obviously a lenient and skewed distribution.

The above results describe only a portion of the data being collected in the project. On-the-job performance data will be collected for the next few months.

My interpretation of the above data and results is that samples are better than forms. The approaches for assessing performance, both past (SAF) and present(interview, role play, assessment center, and job simulation exercise) show a meaningful pattern of results that is generally consistent; in contrast, the typical rating form shows little association.

Assessment for Developmental Purposes. Now I'd like to describe an appraisal program, also for account executives in the same financial institution, composed of candidates who were experienced in financial matters but had limited experience in sales. This group was going to be trained and then moved into the account executive position; no selection or de-selection was considered. The essential purpose of the appraisal process was to assess the KSA's of the group and then use the information gained to design individual, tailorized training programs for each candidate.

Again, the assessment center was used as the appraisal process. The difference here, however, was that the developmental center contained more exercises, 9, as opposed to 5, as well as more complex ones. For example, the selection center's customer walk-in was conducted via telephone in the developmental center as an interruption while the candidate worked on the in-basket. Or, for the "professional prospect" simulation, the client (role player) had more needs and demands, and was eligible for more products.

I do not have the correlational data for this center as I did for the selection center since collection of the same kinds of data is not as necessary here. This was a developmental center. Though discussions were held on the performances observed, there was no need to reach an overall consensus decision; the emphasis was on the dimensions and the strengths and needs of the candidates on those dimensions. Thus, for each candidate a specific training course was suggested.

Summaries of the discussions were prepared for each candidate and feedback to him/her during the course of a feedback interview. Improvement needed areas (development needs) were discussed and classroom training and/or on-the-job training was highlighted. The interview began by asking the candidate to discuss his/her self-ratings on the dimensions and subsequently turned to the assessors' comments. I have no data on the success of the center, except for anecdotal information that was quite positive; candidates appreciated the time given to their assessment, saw the exercises as realistic and relevant, and perceived the entire process as a credible and meaningful one.

Conclusion. My conclusion is that simulated exercises in an assessment center process and a consensual decision making process, can be used to measure *current performance*--which is the purpose of performance appraisals. Such an orientation is consistent with each of the purposes mentioned at the outset of this presentation. For example, assessment centers are obviously useful for (1) feedback and developmental purposes; (2) for promotion and placement purposes; (3) as evidence of content validity in selection criterion and predictor; (4) it should contribute to the development of training curricula; and, finally, (5) development of the exercises as well as the collection of information leading up to the development is, in fact, organizational diagnosis and planning--the exercises can be designed to simulate that which *ought* to be performed.

There are other benefits to samples for performance appraisal. First, we all know of the response biases that emerge when rating scales are used. Assessment centers rely on consensus evaluation; thus, we should obtain fairer appraisals. Second, the assessment is standardized; each candidate is basically placed into the same situation and thus can be readily compared if, in fact, the appraisal decision requires comparisons. Third, it is my contention that supervisors are *not* actually aware of the performances of their subordinates. They have limited opportunity to observe and record behaviors; we all know that the request for an anniversary evaluation is a taxing moment for the supervisor and is replete with all the vagaries assocciated with memory processes. Assessment centers allow us the opportunity to evaluate based on actual performance, recall, recognition, and memory storage issues and problems are not as dominant. Also, whereas supervisors are not thoroughly trained in evaluating others, assessment center assessors are skilled evaluators who devote their attention to assessment and are not concerned with other daily functional duties and performances that are part of a supervisor's job.

I am of the opinion that the situation in which we lack reliable and accurate performance information is a common and frequent occurrence and thus we need an alternative--I recommend the assessment center process as a means for assessing competency of current employees. I am at the point were I am starting to agree with Robert Townsand's (1984) view that "Printed forms for performance appraisals...are used by incompetent bosses in badly managed companies. Real managers manage by frequent eyeball contact." Such contact is by assessment centers and simulations.

SYMPOSIUM

The Necessity for Convergence and Integration of Personnel Sub-Systems

This paper attempts to convey a few simple messages; The first is: If you're not converging and integrating your personnel subsystems--Stop; and do so now! The second is: You probably don't know enough about a job if you haven't seen it performed. The third is: At adding, computers are better than people.

The first portion of the paper is devoted to the second message and reflects experiences of the author with three components: (1) a great deal of time spent watching people work, (2) a little less time, but still a substantial amount, spent hearing people talk about their work (in the sense of describing it), and (3) time spent reading written accounts of what people do on jobs.

Later parts of the paper, again reflecting personal experiences, are concerned with time spent comparing the job analysis records of personnel subsystems at federal, state and local levels, and finding that they were being compared for the first time, even though they had existed in files for years. (The term subsystems here refers to theoretically interacting units or functions, such as recruiting, selection, classification, compensation, performance appraisal, and training. And by job analysis records we mean the complete set of retrievable files that an agency may have in all its subsystems.

Watching People Work. The author comments upon his experiences with job analysis which began in 1949, first with work as a production checker, then as a time and motion analyst using the Bideaux system. That system was used extensively in the textile industry and the experience was in a cotton mill. Observers fairly often attempted to time events three hundredths of a minute in duration and to simultaneously rate those events on scales of both effort and speed.

A great deal of experience was gained looking at jobs in detail. The studies, in most cases, were dealing with jobs for which much of the important work could be directly observed. It wasn't necessary to make many inferences about mental processes such as reading. The point, however, is that they took much more detailed looks at jobs than those typically taken in the "job analysis" context and the process was not a great deal more expensive than those found in public personnel management today. They were concerned with productivity and efficiency and we were making a good profit. In fact, the organization of which I speak was characterized as the most efficient independent metal fabricator in the country during the period in which we were doing this type of job analysis.

Personnel managers in the public sector could also profit by including in their job analyses a larger observation component and depending a little less on subject matter expert recall. My sample of public personnel systems may not be representative, but it does include Federal, state and local merit systems in several federal agencies, in several state agencies, and in several municipalities; and I have not observed much observation on the part of public personnel job analyses. Too many of them do not spend enough time watching others work. Of course the argument for increased observation has been made many times in the job analysis literature. The most recent instance I recall was a strong statement to that effect by Bob Guion at the University of Chicago in 1980. However, as in the case of many other good ideas, there is little evidence of behavior change as a function of those statements. I say confidently that the statements I have seen in several personnel management textbooks to the effect that the observation method is inappropriate for white-collar jobs, are not correct. We have found a great deal of overestimation on the part of subject matter experts of the time they spend reading regulations and other policy documents.

Let's move at this point to the message having to do with converging and integrating your personnel subsystems. I can't find any logical reason for separate, special purpose job analyses as compared to a single, comprehensive analysis such as that proposed by Bemis et al.(1983).

There are three major areas on which this part of the paper is focused. These are (1) problems in subsystem relationships, (2) problems in validity discovery, and (3) problems in validity generalization which are caused by a lack of interaction among subsystems and a lack of a uniform basis for decision making.

It can be argued that trends in the organizational evolution of public and private personnel systems present strong arguments for careful, well-defined, and complete job analysis procedures. The growing complexity of jobs, the sheer numer of jobs and the Affirmative Action/EEO requirements all suggest structured processes for collectin sharing, and utilizing information. The proliferation of computer access and the growing sophistication of data base management software give personnel analysts an opportunity to bring all units in the personnel system into a unified relationship here-to-for impractical if not impossible.

Krzystofiak and Newman (1979) point out that personnel processes such as replacement planning, job rotation, promotions and job evaluations all involve decisions which must be based on job content information which is becoming increasingly complex. Meanwhile, the courts have determined that required validation studies must be based on clearly defined criteria of job performance, requiring, among other things, demonstration of ties existing between identified tasks and the companion knowledges, skills and abilities (KSAs). The integrated job analysis represents a logical basis for dealing with these complex problems.

Many of the problems arising from job complexity and validation requirements are aggravated by the lack of a clearly defined, logical (to say nothing of legal) basis for making decisions and by insufficient or ineffective communications between subsystems.

Consider for a moment four component subsystems in the typical personnel system: recruitment, selection, training and performance appraisal. Historically and naturally the personnel process flows to the right (downstream).

Recruiting Selection Training Performance Appraisal

The Recruiting unit screens on the basis of indicated KSAs, the Selection unit selects on the basis of these KSAs, the Training unit helps employees develop the KSAs into full task performance, and the Appraisal unit evaluates performance of the identified tasks. Clearly this system is based on the necessity to perform certain tasks, thus creating the necessity for task identification and a KSA tie in (i.e., Job Analysis). Once identified, the tasks should determine logically not only the criteria used in performance rating but the KSAs which influence the recruiting and selection processes as well. The foundation, therefore, of the relationship is the job analysis. Any change in the job description should, logically, imply a chain reaction of adjustments in all other related subsystems. From this perspective, the job analysis is not unlike the set of axioms upon which a logician might have based a system of deductive syllogisms. The analysis defines, rightly or wrongly, the body of accepted truth concerning the position. All other actions in the system should be logically predicated upon the existence of this body as truth.

Just as in an axiomatic system, however, the individual properties derived from the axioms-in this case the characteristics of the personnel subsystems and their instruments-do not exist in isolation but are tied irrevocably to each other through the axiom set (the Job Analysis). It is quite appropriate, therefore, for one to require that every action by any of the subsystems be logically traceable to tasks identified in the job analysis.

In this respect, then, a healthy personnel system will also have a flow of information from right to left (upstream).

Recruiting   Selection   Training   Performance Appraisal

Such a flow will serve as a system control indicator. If selection and evaluation are operating in pseudo-isolation, perhaps as a result of inadequate special-purpose job analyses, a situation is created in which subsystems may be working against each other. The source of this problem will be difficult to locate unless there is a routine flow of information in the upstream direction. If, for example, employees are selected on a set of criteria not properly tied to those on which they are evaluated, the training process may be put in a very compromising position. The weaker the connection between selection criteria and performance evaluation, the more lengthy, costly and less effective the training process is likely to be. Thus, to function most effectively, to say nothing of legally, all three subsystems must base actions on the same body of knowledge concerning the job and must communicate with one another.

<u>Effects on Validity</u>. The necessity for evidence of test validity furnishes a sound principal reason for basing the systems recruiting, selection, training, and performance appraisal on a single, integrated job analysis. There are many factors which work to obscure test validity from the personnel analyst (Schmidt, Hunter, Pearlman and Shane, 1979). To allow the problem of validity estimation to be further complicated by a system in which components are based on different job criteria or on different weights for the same criteria seems to be counter-productive if not foolish.

A job analysis is simply a measurement (albeit not entirely quantitative) imposed on a certain theoretical job content domain. Visualizing the job analysis as a measurement and applying the usual logic for analyzing measurements one may examine the potential effects of accurate vs. inaccurate and (2) multiple vs. single job analyses on attempts to discover true validity estimates for selection instruments.

The imposition of the analysis on the content domain will result in some measurement error (E). If E is large, any calculated evidence of validity is likely to be only an indication of internal consistency in the measures. The maximum chance for building a truly valid procedure will occur when E is minimized. The development and implementation of uniformly applied, well-structured job analysis techniques will be necessary to quantify and control E.

In the case of multiple job analyses, which separately serve as bases for the selection and the performance criteria, the internal inconsistency should further obscure the true validity of the selection test.

Generalizability. A long-standing problem in personnel psychology has been that of transportability of valid tests, i.e., validity generalization. Until the last few years the traditional viewpoint has been that test validations are situation specific. It was pointed out in 1976 by Guion that the inability to generalize validities makes it impossible to develop the general principles and theories that are necessary to take the field beyond a mere technology to the status of science.

Promising results began to emerge in this area, however, when Schmidt and Hunter (1977) and later Schmidt, Hunter, Pearlman and Shane (1979), using a Bayesian model, produced evidence that between-study variations in observed validity coefficients are, at least in some cases, artificial in nature. If, indeed, validity is reasonably stable across time and situations for similar jobs, one could apply validated selection tests to new situations without carrying on the usual validation study. The potential savings in time and money are obvious. The only thing that would be necessary in the new situation would be a job analysis in order to insure that the job at hand was in fact a member of the class for which the test was validated. There are at present, however, many obstacles to validity generalization. Schmidt et at. (1979) pointed out that validity studies must be more complete than they have typically been in the past. A common problem has been the incomplete identification of the job being studied.

Associated with the problem of validity generalization are several questions which have a quantitative flavor. First, how "similar" must the job analysis in the new situation be to the old in order to justify transportability of the test? This poses a simple dichotomous choice which one may resolve by standard statistical procedures; but still unanswered are the questions which seek to quantify the effects of job differences on validity. For example, can a function be constructed that describes validity as a function of similarity? To do this it will be necessary to identify the error variance in the job analysis procedures or to show that its effects are minimal. Generalization will likely involve a movement away from subjective job ranking methods to a well-structured uniformly adopted process for collecting and utilizing job information. The personnel system made of interacting subsystems using a common data base with an integrated, self-sharpening job analysis as its foundation seems to be the logical model for extending research, maximizing effectiveness and minimizing cost.

Economic and Legal Arguments. Does the building of such a system make economic sense? There are data which would lead one to think so. Let's look at a summary of some recent EEOC data concerning monetary benefits for victims of employment discrimination. For fiscal years 1980 through 1983, awards under Title VII were, respectively, $13 million, $9.8 million, and $13.5 million. Under ADEA, for the same years, they ran 2.3, 1.3, 20.5, and 24.7 million dollars. Under EPA awards (in millions) were $1.8, 3.2, and 2.0 for the last 3 years. The grand total of this is something over $110 million, so we're in a fairly large economic ballpark. Obviously, not all of these dollars can be attributed to job analysis problems, but, surely, a few of them can be. There is the citation of the case, in which a court "barred the further use of a civil service test developed at a cost of $1.25 million due to improper validation procedures". The problems in that case centered around job analysis.

Implementing an Integrated Approach. Can one implement a job analysis system which will provide data for all personnel functions? Returning to Bemis (1983) for a quote, "It should be clear that there is no one right or best job analysis method, although in presenting the Versatile Job Analysis System, the authors express their preference for a comprehensive approach which develops a data base that can be applied to the full range of personnel functions." Bemis goes on to describe the applicability of what he terms a versatile job analysis system (acronym VERJAS), which includes job design, classification and evaluation, recruitment, selection, training and performance appraisal. We recommend it to all not already familiar with it.

We also recommend Schwartz' chapter in Bemis et al.(1983). The portion of Schwartz' chapter dealing with the role of the computer in the automation of personnel systems is particularly enlightening. Schwartz gives examples of how a computerized data base for job analyses could effectively contribute to personnel management: "Managers could key in any changes in positions or equipment as they occur; classifiers could request a current listing of all tasks, responsibilities, and skill requirements; recruiters could request a current listing of the three most important duties, and any unusual context factors; staffers could request a listing of basic and special competencies not learned on the job; trainers could request a listing of all positions where operation of a specific type of equipment is required; validation experts could request a listing of all positions where certain specific tasks are performed; and managers could cross reference to tasks in existing positions rather than write up new task descriptions."

Now to summarize: (1) When you're doing a job analysis, be sure you have included an adequate observation period, even for white collar jobs. (2) If the subsystems of your personnel system do perform separate job analyses, be sure that they have a very good reason for it. It's highly likely that they're not only reinventing the wheel, but that the new wheel is a different size from the old one and will make the ride through litigation very bumpy. (3) If at all possible, do one good job analysis, (perhaps a la Bemis) for the whole system, and automate the storage of job analysis data so it can be kept up to date.

PAPER SESSION

Cost Effective Measures

Chair: Stephen Boles, San Mateo County Personnel Department

Predicting Test Performance: A Content Valid Approach To Screening Applicants

Ronald D. Pannone, The Port Authority of New York and New Jersey

Since applicant screening decisions are related to selection decisions, screening decisions should be made on the basis of job relatedness,.both logically and statistically. Screening criteria have one common objective: to improve the cost-effectiveness of testing programs by eliminating the grossly unqualified before participation in costly testing programs.

A review of the literature demonstrate that biographical questionnaires are effective predictors of a variety of criteria. Biographical questionaires provide a standardized approach to evaluating applicants' backgrounds, are easily administered and scored, and highly cost effective when utilized in employement settings.

Mosel (1952) classified biographicalquestionnaires as empirical or rational. Empirical questionnaires are usually developed according to criterion related validity models. Scoring weights are given to the test items based on the magnitude of the relationship between the item and the criterion. These weights are to eliminate low criterion subjects but not high criterion subjects. Empirically developed biographical questionnaires have been found to be effective predictors of performance on aptitude tests (Sparks, 1971) and assessment center performance (Ritchie & Boehm, 1977 and Quaintence, 1981).

Content validity models are the basis for rational biographical questionnaires. Questions focus on previous work experience that relates to the requirements of a given job. They have predetermined standards by which raters judge che applicant's responses.

Method. This study developed a rational biographical questionnaire that would predict content valid test performance for electrician applicants. The need for this arose because of a consistently low pass rate on the electrician's evaluation. There was a large number of applicants and very few received a passing score. The study contended that previous work experience in terms of a specific domain is quantifiable, and will predict performance on content valid tests designed to reflect that domain.

The biographical questionnaire was developed along the ideas proposed by Wernimont and Campbell (1968), the behavioral constency model; "the best indicator of future performance is past performance", and Asher and Sciarrino (1974) who contend, ' . . . the more points in common between the predictor and criterion space, the higher the validity coefficient. . .".

These models suggest (in terms of study), a biographical questionnaire made of these items that are behaviorally consistent and capable of demonstrating a point-to-point correspondence with the criterion will have a higher validity coefficient than the traditional screening criteria of years of training and years of experience.

Task statements were generated for the questionnaire and the applicants were required to rate their previous work experience with regard to each task on a four-point scale. The scale evaluates the level at which the applicant did or didn't perform the task: (a) previous job(s) didn't require me to perform this task, (b) I performed this task under direct supervision, (c) I performed this task independently, and (d) I supervise(d) others performing this task. There was also an unscored "fake" item which described a non-existent task. This was to check the amount of faking that may occur.

Results. The reliability of the questionnaire, computed with coefficient alpha, was .96. The split-half reliability was .86. Both indicate a highly reliable questionnaire.

The questionnaire scores showed a stronger relationship with the criterion than did the education and experience requirements. The biographical questionnaire correlated significantly with the written test (which asked objective "electrician's" questions), $r=.42$, $p.<.01$ and years of work experience, $r=.30$, $p<.01$; $n=221$.

There were 45 people who passed the test and 176 failures. The difference between the means of these two groups was highly significant, ($p<.0001$) and there was a low overlap between the two distributions.

The "fakers" were those who had answered positively on the fabricated task. The "non-fakers" answered they had not done this task. The non-fakers comprised 65.2% of the group and their responses (on the biological questionnaire) showed a stronger relationship with the criterion than did years of work experience and years of electrical training. There was a .55 correlation ($p<.01$) between the biographical questionnaire and the written test. The fakers' biographical questionnaire score correlated significantly with the written test, (although not as strongly as the non-fakers), $r=.26$, $p<.01$, $n=75$. The non-fakers had significant correlations between the biographical questionnaire and work experience, $r=.35$, $p<.01$; the written test and years of electrical training, $r=.24$, $p<.01$; and the written test and years of work experience. None of these correlations showed up in the fakers category. The difference between the means for the fakers and the non-fakers was significant for biographical scores ($p<.0001$) and years of experience in electrical maintenance ($p<.01$).

Coefficients of .42, .55 and .26 between questionnaire scores and test scores for all subjects; fakers, and non-fakers, respectively suggest that rational biographical questionnaires developed by the techniques used in this study are valuable in screening applicants. Using this kind of specific biographical data as screening is superior over broad screening criteria such as level of education and/or years of work experience.

The study here also shows that the information sought in the questionnaire is highly susceptible to faking. It is clear that this faking, as evidenced by one fake item, introduced error variance that distorted the validity coefficients.

Future research should examine the effectiveness of scoring systems based on task importance ratings and more fully explore the effects of falsification on validity coefficients by including a scale of items to detect faking.

Reducing the Size of a Candidate Group By Providing Feedback on Selection Probability

William E. Donnoe, California State Personnel Board

Recently, it has been common to see thousands of people make application for entry level examinations. Without the knowledge that they are competing with so many others, applicants proceed through the examination process. The successful candidates attain ranking on an employment list, yet their probability of being hired is slim. By informing candidates of their probability of success in an examination, it was believed that the number of candidates self selecting out of the examination could be increased.

In a recent administration of a California State civil service, entry level clerical examination, ne-half of the applicants were given information indicating the total number of applications accepted for the examination and the number of expected hires to be made from the employment list. Those applicants receiving this supplemental information were randomly selected from the total group. Comparisons in drop-off rates, sex ethnicity, and test scores were made between the group receiving this supplemental information and the group that did not. The notice-to-appear for the written test was used to deliver the supplemental information to the applicants. The content of the supplemental information was as follows:

THE STATE PERSONNEL BOARD HAS ACCEPTED 5775 APPLICATIONS FOR THE OFFICE ASSISTANT EXAMINATION. IT IS ANTICIPATED THAT THE EMPLOYMENT LIST GENERATED BY THIS EXAMINATION WILL RESULT IN APPROXIMATELY 110 HIRES PER YEAR.

Following administration of the written test, results were tabulated on the overall effect of the information including estimates of cost savings, the effect of the supplemental information on different subgroups of applicants (by both sex and ethnicity) and the effect of the information on candidate abilities (as reflected by written test performance).

The results of this project indicate that the group receiving the supplemental information had a significantly higher drop out rate than those who were not informed of their probability of success (Using the test for significance between two proportions, a 'Z' value of 2.64 was observed: $p < .01$. The estimate of cost savings for this one examination was in excess of $3,500.00. This estimate is based on the reduction in the number of scheduled interviews which followed the written test for successful applicants.

The ethnic and sex composition of the two groups was very similar (due to the random assignment of applicants to groups). Of those applicants who self selected out, no significant differences were observed across sex or ethnicity. This indicates that all subgroups perceived the information consistently.

The written test scores of the two groups were compared to determine if a relationship existed between test taking ability (as measured by the written test used) and perception of the supplemental information. The results indicate that no differences exist between the two groups on their test scores. This further evidence that the use of supplemental information intended to encourage applicants to self select out of this examination, functioned on a random selection basis and was perceived congruently by different groups of applicants.

The discussion of this procedure to reduce candidate groups will include cost savings possibilities for agencies, and implications under the Uniform Guidelines.

Validity with Economics: Techniques for Evaluating 9,000 Applicants

Roger Davis, King County Personnel Department, Seattle, Washington

Coping with large numbers of job applicants requires not only some validity in evaluation techniques but adequate budget. The irony is that in an era of cutback management, personnel organizations usually find themselves in the bind of extra economic constraints and extra job applicants to process.

Typical organizational responses to this problem, whether tinkering with minimum qualification standards to raise or narrow them, re-using worn-out multiple-choice tests, or borrowing another jurisdiction's test items, are typically risky, unsatisfactory, and more importantly almost always inadequate, professionally, legally, and technically. Those old tricks are only adequate at best economically--or rather just in terms of front-end costs.

An alternative technique that would contribute toward resolving this issue would be to develop a simple, cheap-to-reproduce, non-secure test instrument that would allow applicants to compare themselves directly to the requirements of the job in terms of functional criteria of performance as well as other actual hiring standards.

Of course to do that, you have to develop, hold, and maintain the standards--which is not, for most organizations, an easy thing to do. We did this in King County (Seattle, Washington) to hire Police Officers, and have been doing it since 1981 with positive results.

There is the statement that "There is no such thing as a free lunch." That is a notion I tend to agree with, although I imagine that there are many legitimate ways that a good lunch can be earned. In the world of work the notion of no free lunch translates into the concept that there is a cost associated with most every activity, and when we are productive, a benefit as well, a "return-on-investment" as it were.

One of the problems in personnel management is that by and large we do a very poor job of identifying the costs of our activities, planned and actual, let alone the dollar benefits of our activities.

Cost accounting in personnel management and testing is only one side of the dollar equation. The other half is benefit accounting. And here, in evaluating in dollars the benefits of our products and services we in personnel management tend to be just no good at all. We don't know the value of what we do. That is probably because we don't calculate it, and that in turn is probably because we don't know how to calculate the beneficial value of what we do--because we haven't learned how yet.

What I am really talking about is not the economic value of your overall services, but that of the specific services or products you contribute. Test utility analysis (reach for your Taylor-Russell tables, please) contributes valuably to our understanding the overall organization's economic gain through our selection projects. Other activities need to be seen as similarly "gainful employment" too. Validation is a gainful activity, and can be understood as such by top management if you don't try to sell it as an activity independent of test selection, development, and administration. When it is separated, validation sounds too much like "research." Policymakers don't like to buy much research.

Aside from gains, another economic benefit of a personnel activity may be a savings. Some things we do in personnel generate a specific, definable dollar savings. It is worth measuring that.

Savings usually means doing something less than what you formerly did. You cut something out to make a savings. You're doing less, and being cheaper thereby. But that isn't always the case. Sometimes you can do more, do something extra, and effect a savings. Which brings me to the topic of this paper.

In 1981 I added a measure to the battery of serial tests comprising the Police Officer examination for King County. The purpose of this step was paradoxically to do less testing. The added measure was a simple, inexpensive one which was intended to and does serve as an applicant population-management instrument. It reduced the applicant population of unqualified candidates so that fewer people were processed in subsequent, more expensive testing.

The device added is a checklist of all the principal employment criteria that we can identify, other than the implicit performance standards in the competitive tests in the examination.

Candidates are invited to compare their own qualifications against these criteria, and if they are not met the applicant sees for himself immediately and early that he does not meet our standards, does not qualify to become a police officer in our agency at this time, and will not be hired. It saves everyone's time, energy, and expense.

To briefly describe the instrument, it is in a word a test--not perhaps in the technical-professional sense of that term, but in the legal sense. This constitutes a test.

To describe the test somewhat more fully, in a phrase, it is a self-scored, non-weighted, content-validated criterion sample.

It is a set of hiring standards.

It is the answers to the test, used as the test itself without the questions.

It is the secrets, in this case, to getting the job. If you have these qualifications--in some dimensions, the more the better--the job is virtually yours.

As a set of employment standards they are obviously not arbitrary. As soon as you read them you see they are quite functional. The checklist consists of eight legal and regulatory requirements drawn from statutes and rules, eight key medical standards which were drawn from a comprehensive, special list, then some background investigation standards, and some 65 performance standards.

The 65 performance criteria were derived from a job analysis conducted by my predecessor and the U.S. Civil Service Commission's regional psychologist in 1977. Following the Job Element method published by Primoff (1974), the job analysts had broken down the work of a King County Police Officer into more than 600 job elements, with ratings from a panel of superior officers and supervisors.

Unfortunately, once they got all this good data the job analysts were not quite sure what to do with it, and consequently did little. After I came on the scene in 1979 I had a personnel analyst go over the job analysis results and, in job element language, draw out the items high in Barely Acceptable, low in Superior, and high in Trouble Likely. If you are not familiar with job element terminology, it means that we were looking for performance standards to serve as minimum qualifications. Such items simply screen applicants in or out of a competition.

In this way the Checklist was formed to meet the determined purpose.

What are the savings benefits of introducing this instrument in the manner described?

Prior to introducing this device to our examination we had experienced in 1979 and 1980 No-show rates (failure to appear for testing) of 14% and 15%. These were declines from the original numbers of applicants each year to the number of candidates appearing for the first test.

After introducing this instrument in 1981 and 1982, these rates increased to 31% and 26% respectively. The net effect, we have conservatively calculated, was a 12% drop in our favor of candidates whom we otherwise would have had to process with attendant cost and time burdens.

If an agency were testing, even over several years, say 9,000 applicants (or 8900 as we could forsee) using this Checklist would have saved having to process more than 1,000 unqualified persons.

If an agency were using a cheap first-stage test, for example one that cost $2 per capita, the savings by the use of this Checklist would have been about $2,000--the cost of bringing two staff members to the IPMAAC meeting. If the agency were using a more expensive test, such as the new IPMA or ETS police or fire tests (as King County does) at $6.50 to $8.00 per capita, the savings would have been about $7,000 to $8,000. There is a large amount of research or professional development that can be bought for $7,000-$8,000. When you crank in overhead, miscellaneous, and related costs that are saved by having 1,000 fewer people to test, or 12% of your applicant population, an even greater savings can be realized.

This then is what Validity with Economics means. Demonstrable savings is one of the definable economic benefits that yield from testing and validation. Look at it this way. testing and validation, in private-sector parlance, can be seen not as a cost center but a profit center.

Professional Affairs Committee Presentation

Ethical Issues Involved in the Use of Polygraph Tests in Selection

Patrick T. Maher, Personnel and Organization Development Consultants, Inc.

LaPalma, California

This paper presents some of the ethical issues inherent in the use of the polygraph for employment purposes. Significant controversy surrounds the use of the polygraph as a basis for making employment decisions, even in such sensitive occupations as law enforcement. Employers justify the use of polygraphs because of the need for honest employees. Others criticize its use because they feel the polygraph is inaccurate and its use constitutes an invasions of privacy.

Legal Considerations

Ethical Issue Involved. When specific laws do not exist, how does the employer determine the appropriate uses and limits of polygraph examinations in the selection process?

Discussion. The federal government and 33 states do not prohibit the use of polygraphs by an employer, either as a pre-selection device, as part of a routine, periodic test of continuing honesty among employees, or as the means of identifying guilt or innocence of specific employees involving a specific incident. There are no legal guidelines concerning the uses and limits of the polygraph by employers, except in states that have prohibited the use of the polygraphs and from court decisions where the polygraph may have been inappropriately used in specific cases.

Accuracy-Ethical Issue Involved: Is there justification for using the polygraph as the basis for selection when potentially the greater number of persons identified as being dishonest are in fact honest?

Discussion: There is debate as to the accuracy of the polygraph as a "lie detector". Supporters claim 90% accuracy. Other researchers, however, disagree, with studies indicating accuracy as low as "just better than chance." Nearly all studies claim that the accuracy rate is higher with test results showing a subject is truthful than with results showing deception. Critics argue that many things can affect the results of a polygraph, therefore, lowering its accuracy. Breathing, slight movement, reaction to certain sensitive areas, stress, and other psychological reactions can affect one or more of the physiological reactions being measured.

In deciding whether or not to use a polygraph and how it is to be used by employers, it is mandatory to consider the following question: "What are the consequences of using a polygraph for selection purposes?" Under the assumptions that the polygraph is 90% accurate, with 95% of all candidates truthful and 5% of all candidates liars, from a sample of 1000 candidates, 950 would be truthful, and 50 would be liars. If the rates of accuracy of the polygraph is 90%, 45 liars and 855 truthful candidates would be correctly identified. On the other hand, the polygraph will classify 140 as liars. Forty-five will be actual liars, and ninety-five will be truthtellers falsely identified as liars. Therefore, out of the 140 liars, 68% are telling the truth. If 500,000 to 1 million employees or potential employees take polygraph examinations annually, a 90% accuracy rate will incorrectly identify 50,000 to 100,000 employees annually. Even assuming that only half of those are being truthful, 25,000 to 50,000 honest employees will be improperly labeled as liars.

Test Results. Ethical Issue Involved: How does the employer deal with inconclusive test?

Discussion: A polygraph machine measures involuntary physical responses that are triggered when a person is lying. Those responses include a faster heart beat, changes in blood pressure, and increased respiration. The polygraph operator must take the quantitative information that is recorded and subjectively interpret it to determine honesty. Thus, the polygraph examination is really a two-part process. First the machine records physiological responses to questions. Then, the examiner must decide what those responses mean.

The examiner must make one of three determinations: that the response (or overall test) indicates that the candidate is being truthful, is not being truthful, or that the test is inconclusive. That is, it is not possible to determine that the candidate is either truthful or not truthful.

Qualifications of Examiners. Ethical Issue Involved: What obligation does the employer have in ensuring that the polygraph examiner being used is a qualified examiner and how is it determined if he or she is qualified if there is no licensing requirement in the state?

Discussion: Although great emphasis is placed on training, few states have any means for licensing or otherwise determining the fitness of polygraph examiners. Furthermore, there is little or no control over schools or courses that allegedly train and certify polygraph examiners.

Examination Questions. Ethical Issue Involved: How does the employer ensure that questions asked during the polygraph examination are limited strictly to issues that are clearly demonstrated to be job related, especially when it is conducted by someone other than an employee of the agency.

Discussion: Examination questions for selection purposes are usually restricted to those that are job-related. Often though, the polygraph examiner is not limited in this way and questions involving personal honesty and habits such as "With whom do you live?" and "Do you drink?" are asked. Problems also exist when the questions and answers are not used directly by the employer to make the employement decision, but instead the employer only uses the examiner's interpretation of the answers.

Polygraph Interviews. Ethical Issue Involved: Is it ethical for an employer to use the polygraph as a psychological procedure to force damaging admissions as opposed to using it as a scientific procedure to detect deception?

Discussion: Pre-employment polygraph tests are frequently preceded by an extensive interview. This is where most damaging information comes out rather than in the polygraph exam itself. Through various "tricks", the perspective employees are led to believe that the polygraph is infallible. They will frequently admit damaging information in the belief that the machine will detect it in any event.

Policy on the Use of the Polygraph. Ethical Issue Involved: What is a proper policy on the use of polygraphs in the selection process?

Discussion: The following suggestions are made as to "proper policy": The terms "pass" and "fail" should not be used in discussing polygraph results. Rather, the results should be: (a) No questionable responses: (b) Questionable responses; (c) Applicant has admitted the following information.

Polygraph results should not, by themselves, be used to disqualify an applicant. Questionable responses should be verified by an admission by the applicant or by independent background investigation.

Only trained, experienced, qualified examiners should be used.

The applicant should be advised at the start of the application process that he or sle will be required to take a polygraph examination.

Polygraph examination areas and individual questions (including control questions) should be limited to those that are clearly related to the target job.

Results of polygraph examinations should be carefully controlled with severely restricted access.

Once the applicant is hired, polygraph examination results should be destroyed or sealed.

If an applicant is hired, polygraph examination results should be sealed and not made available except for future employment issues (e.g., civil service appeal or future application) involving the agency that conducted the examination.)

The polygraph examination should be administered to all applicants entering a background phase, not just a select few.

References

Andrews, Lori B., The Employer's Lie Detector: How Fair Is It? "The Los Angeles Daily Journal," July 29, 1983, p. 5.

Elmore, Richard A., The Polygraph: Perceiving or Deceiving Us? 13 N.C. Cent. Law Journal 84-100, Fall, 1981.

Polygraph Test not Discriminatory -- Arrest Inquiries Are. Decision of EEOC --Decision #76-12, August 15, 1975.

Regulation of Polygraph Testing in the Employment Context: Suggested Statutory Control on Test Use and Examiner Competence. 15 UCD Law Review .13-133, Fall, 1981.

Stein, James L. and Barbara D. Dennis, eds., Truth, Lie Detectors, and Other Problems in Labor Arbitration. "Proceedings of the Thirty-First Annual Meeting of National Academy of Arbitrators, April 4-7, 1978. Washington, D.C.: Bureau of National Affairs, 1979.

SYMPOSIUM
Organizational Change

Organizational Innovation

Marianne Bays, Prudential Insurance Company of America

My review of organizational innovation research began with the question of what makes one organization more prone to innovation than another.

I found that the organizational innovation research falls into four distinct subcategories:

The 1st is: studies that address the attitudes, personality or other individual characteristics of those who innovate.

The 2nd is: studies that focus on the manner in which organizations acquire the information that they need to keep up with technology advances and to remain innovative.

The 3rd category includes: studies that try to isolate individual, organizational and environmental variables which affect organizational innovation adoption and diffusion behavior.

The final group includes: studies that look more broadly at the process of organizational innovation and seek to understand it.

Individual Characteristics of Innovators. If some individuals are more likely to innovate than others and the individual differences between more or less innovative persons can be identified and measured, then organizations might be able to use this information fruitfully in their staffing decisions. This thought is one that has spurred a distinct line of research in the organizational innovation literature.

Michael Kirton is one of the most active researchers in this area. He hypothesized that an innovative orientation (that is the tendency toward "doing things differently") was an extreme of a cognitive personality dimension the other end of which was an adaptive orientation (or tendency toward "doing things better" within existing structure). He further hypothesized that all people could be located along an adaptive to innovative orientation continuum and that adaptors and innovators would bring incommensurable viewpoints and different solutions to administrative and organizational problems.

Kirton's research suggested that adaptors would be methodical and conforming and at home in a bureaucracy because of their problem solving approach. Innovators on the other hand, were conceptualized as organizational loners who have little awe for traditions and almost compulsively try new approaches.

Kirton developed and tested a self-description instrument that could be used to classify individuals along an adaptive to innovative orientation continuum. In repeated uses of this instrument, it was found to yield results that point to the even distribution of adaptors and innovators in the population as a whole.

Kirton attempted to extend the research on innovators in another study which looked at the proportion of innovators vs. adaptors in different types of organizations. He categorized departments within one organization as either primarily concerned with their own internal processes or acting as interfaces between other departments or between the company and the outside world. He hypothesized that the first category of departments would have a lower proportion of adaptors within them than the second category of departments would. He also posited that departments with higher average innovator scores would have a more turbulent environment to deal with than departments that were more adaptor weighted. Support was also found for the hypothesis that the innovator-oriented departments would have a more turbulent environment.

Kirton's research and that of others has shown that some individuals do seem to be more prone to innovative attitudes, values and, perhaps, behaviors than others. It has also shown that organizations seem to have a mix of innovative-oriented employees and adaptive-oriented employees which differs by their task needs. Kirton's work also suggests strongly that those of us who work to implement organizational change face a natural acceptance barrier from those in the organization who have a more adaptive, less innovating orientation to problem solving.

Organizational Communication Behaviors and Innovation: A second line of research which is pertinent to the question of what makes one organization innovate more readily than another is that which focuses on organizational communication behaviors. I highlight some of the research: 1. A number of researchers have found evidence of special boundary roles that exist to link innovating organizations to their external environment. Among these roles are: Internal Communication "Stars" or Technology "Gatekeepers"- these people serve as communication network nodes by conveying information from external domains into the internal communication network of the organization. There are also what are called: "Organizational or Laboratory Liaisons"- these people serve in an internal communication role which works to link parts of the organization together. 2. It has also been found that the greater the work related uncertainty or task interdependency in the organization, the more special boundary roles are needed to deal with the uncertainty and that either too few boundary roles or too much redundancy in boundary roles can result in lower organizational performance.

Determinants of Organizational Innovation: Unfortunately, the studies that were taken to address these problems were so full of methodological problems themselves that they really did not accomplish their goal of extending knowledge in this area. In sum, we know very little about the combined and separate effects of individual, organizational and environmental variables on organizational adoption of innovation.

The Process of Organizational Innovation: The question of how innovations are diffused within and across organizations has stimulated a more fruitful line of research in the organizational innovation literature. Researchers have tended to take a case study approach or a historical perspective across a set of organizations in this work.

Richard Walton studied eight organizations that made early (beginning in the 1960's) and initially successful innovations in the area of comprehensive redesign of work. His focus was on identifying how much diffusion occurred of the innovation within these firms, what the vehicles for diffusion were, what barriers to diffusion of innovation were encountered and how the character of the innovation affected the rate of its diffusion.

Differences and similarities among the sample companies were explored for evidence of the factors which helped or hindered the innovation diffusion process.

Findings were as follows:

1. The organizational support for diffusion of innovations can be impeded by emergent weaknesses in pilot projects over time.

2. A pilot project, even if it continues successfully, can impede diffusion of the innovation if it lacks visability or credibility in the rest of the organization.

3. Management's ineffective formulation and communication of the diffusion policy can discourage diffusion.

4. Inadequate follow-through in terms of locating accountability for accomplishing the diffusion of the innovation or providing "how to" knowledge can impede progress.

5. Top management commitment is needed to achieve diffusion of innovations.

6. Vested interests on the part of unions, bureaucrats, and other affected employee sub-groups can impede _or_ aid diffusion progress.

7. Walton's final finding was that pilot projects may experience self-limiting dynamics which damage the innovation diffusion progress.

Walton's findings may have limited applicability to the diffusion process for organizational innovations other than work restructuring. Recognizing this, he has attempted to construct a framework for comparing innovations in terms of how easy or difficult they are to diffuse. These are the factors that are suggested for such comparisons:

1. _Relative Advantage_ - How easy is it to cost justify the innovation?

2. _Communicability_ - How straight-forward and readily grasped are the proposed changes?

3. _Compatibility_ - How congruent is the innovation with existing norms, values and structures?

4. _Pervasiveness_ - How widespread are the required changes in terms of their impact on the existing organizational system?

5. _Reversibility_ - What are the consequences (costs) of reversing the innovation? Can status quo be easily restored?

6. _Number of Approval Points_ - How many approval channels must be satis-

fied before the innovation can be adopted?

7. Transportability - Can the innovation be adopted "as is" or must it be "tailored" to fit each new unit in which it is implemented?

Using these criteria, work restructuring can be seen to be a particularly difficult innovation to diffuse. As a result, Walton's study, despite its lack of control and its reliance on after-the-fact data collection, has yielded a wealth of information about the problems that can be encountered in the innovation diffusion process. It has, at least, set the stage for future research identifying variables that seem to impact on diffusion success. It may also have provided some useful advice t· practitioners.

To conclude, my review of the organizational innovation literature failed to identify much consistency in the approaches taken to study of the subject or in the research findings themselves. A number of serious methodological problems in this literature were identified as well. At the risk of sounding trite, there is a need for further integrating research in this field. At least, some of the significant findings reported by researchers of aspects of organizational innovation could benefit from cross validation on different samples drawn from different types of organizatons and on different types of innovations. However, not all of the independent variables used in previous research are, in my opinion, equally worth pursuing. Further, while several researchers have posited the existence of innovation characteristics that act as moderating variables, there remains a need to test these more carefully.

Western Region Intergovernmental Personnel Assessment Council (WRIPAC)

Invited Speaker

Contributions of Personnel Professionals to the Bottom Line

Dr. Wayne F. Cascio, Graduate School of Business Administration
University of Colorado at Denver

For quite some time now I have had the uneasy feeling that much of what we do in personnel/human resource management is largely misunderstood and underestimated by the organizations or agencies we serve. At least in part, I believe, we ourselves are responsible for this state of affairs since much of what we do is evaluated only in statistical or behavioral terms. Like it or not, the language of business is dollars, not correlation coefficients. The utility, or payoff to the organization, of the marketing department, the accounting department, even the company lawyer are not questioned, because a monetary referent to the value of their services is readily apparent. If the organization had to contract out for their services it might well cost far more than the cost of retraining them on a full-time basis. In addition, the magnitude of dollars saved through skillful legal maneuvering to avoid a lawsuit, or to win one if a suit is unavoidable, or the dollar value of an effective advertising campagin, can be estimated in a fairly straightforward manner by top management.

Unfortunately, in many situations there is no such yardstick of the relative worth of the personnel professional. The implicit assumption in our work, of course, is that statistically significant results yielded by some time of assessment procedure imply large amounts of dollars saved. While this is true, the actual size of the net dollar savings to the firm often remains unknown and unverifiable. This need not be the case, for the technology is now available to demonstrate the dollar value of all personnel/ human resource activities, not just some of them. In my opinion, this is an exciting time for assessment specialists, for we are on the verge of a marked shift in the way we market our services and in the way we demonstrate the value of our research.

In a moment we'll consider some familiar examples, but first it is important to dispel some popular misperceptions about our approach. Our approach is not "Human Resource Accounting" (HRA), it is "Behavior Costing." We are not simply tabulating the sum total of a firm's investments in its employees, or discounting the expected value of their future earnings, as HRA suggests. Instead we are placing dollar values on the economic consequences of employee behaviors such as absenteeism and turnover. It is these economic consequences that lead to large costs (or cost savings) for firms, and this approach dovetails nicely with the kinds of activities that many personnel professionals are involved in. Let me emphasize one further point: behavior costing does not imply that we need to trade in our measurement tools and research designs. It does not imply substantial retraining in accounting or economics. But it does require a break with tradition, for what is required is that we carry the results of our research as Personnel Professionals one step further in order to translate them into estimates of dollars gained or saved. Here are some familiar examples.

Absenteeism can be defined as any failure to report for work, as scheduled, regardless of reason. The literature suggests several approaches for dealing with this problem: job enrichment, point systems, poker hands, and OB Mod, just to name a few. We can bring our training and expertise in measurement to bear on this issue, for in considering the drop in absenteeism from Year 1 to Year 2, savings of $500-$1,000 per employee are not unusual. In fact, one Canadian industrial/organizational psychologist I spoke to recently told me that his 30,000-employee bank saved $7 million in one year simply by installing a computerized absenteeism reporting system! The very act of measuring absenteeism, coupled with measurement of the per employee cost of absenteeism, made such an estimate possible.

Turnover may be defined as any permanent departure beyond organizational boundaries. When the total cost of employee turnover (i.e., separation costs, replacement costs, and training costs) is considered, less the positive value or gains to the firm associated with those who leave, the net aggregate cost can be substantial. At a major brokerage house that I studied using the methods described above, total annual turnover costs were $1.5 million, with a cost per terminating employee of almost $7,000. Yet, if through proven personnel techniques such as realistic job previews , work redesign, or performance testing, voluntary dysfunctional turnover could be cut by just 10% annually, it would save the company over $100,000 per year. And that's called, "Making ourselves useful".

Job attitudes, particularly job satisfaction regarding pay, promotions, coworkers, supervision, and the work itself, are directly related to variables such as tardiness, turnover, absenteeism, strikes, and grievances. They are related indirectly to job performance, for if a dissatisfied worker stays home instead of coming to work, the productive value of his or her labor (and its associated dollar value) is lost. Methods for relating changes in attitudes, reliably measured, from Time 1 to Time 2, to the unit costs of employee behaviors, were proposed by Likert and Bowers in 1973, and refined by Mirvis and Lawler in 1977. In the latter study, using 160 bank tellers drawn from 20 branches over a one-year period, Mirvis and Lawler evaluated changes in attitudes regarding intrinsic satisfaction, organizational involvement, and intrinsic motivation, and related the changes to the unit costs of employee absenteeism, employee turnover, and balancing shortages. Results showed that a one-half SD improvement in job attitudes yielded a direct savings of $17,600 per year and an estimated total savings of $125,000 per year (in 1977 dollars). Results like this, when validated to show the actual dollar savings obtained, have a powerful effect on top management.

Personnel selection and training. Using the linear regression-based, general utility model of Brogden-Cronbach-Gleser, assessment specialists are now able to estimate the net dollar benefits of valid selection and training programs. The big stumbling block, of course, is the estimation of the standard deviation of job performance in dollars. Previously it was thought that cumbersome cost accounting procedures or industrial engineering-based work measurement methods were required to estimate this parameter. But advances in research, specifically the Schmidt-Hunter global estimation procedure, the 40% rule, and the Cascio-Ramos Estimate of Performance in Dollars (CREPID), have reduced this problem. Research on the comparison and validation of these methods is a required next step, and it is presently underway.

In the selection of just 10 claims approvers per year at Life of Georgia, for example, it was demonstrated (through actual work measurement methods) that the use of a test with a marginal validity of .22 over the previous procedure, yielded an annual dollar gain in productivity of $30,000 to the company.

A similar approach can be used to evaluate the dollar gains associated with training and other organizational interventions, except that the effect size d (the difference between the means of the experimental and control groups on some treatment, expressed in standard deviation units) is substituted for the validity coefficient.

As for future research needs in this area, it seems to me that what we need are meta-analyses of the impact of properly conducted organizational interventions (e.g., team building, goal setting, work redesign, assessment centers) expressed in dollar terms. I say "properly conducted organizational interventions", because a danger in the blind application of meta analysis techniques is that we may be making inappropriate generalizations if the research in question was sloppy or poorly conducted to begin with. All of us have the background and training to do sound experimental research. Now we also have the technology to translate our results from behavioral or statistical terms into the language of business (i.e., dollars). Isn't it time we got on with the job?

INVITED ADDRESS:

"Comparable Worth"

Dr. Helen Remick, University of Washington

The issue of comparable worth is both a political and social movement according to psychologist Dr. Helen Remick. Globally, comparable worth signifies economic equality and economic power which have been denied to women. More specifically it deals with job evaluation of all jobs. The traditional job evaluation which deals with, for example, only professional or clerical jobs must be made a thing of the past.

Comparable worth differs from traditional job evaluation in four major ways. First, comparable worth covers all jobs. Thus, it has a broad scope. Second is the idea of intent. Comparable worth is used to justify the existing salary structures. Third is the equity issue which involves evaluating female dominated jobs. Last is the sensitivity to issues of sex bias in job evaluation. Comparable worth is meant to be sensitive to sex bias.

Comparable worth requires input from both employees and the public in order to achieve a broader sense of direction. Good employee input is a must.

Dr. Remick brings out the point that the public sector has been involved in many studies involving comparable worth and she questions why this is so. She answers this in stating that the public sector wages are made public and information is more accessible. Also she points out that the public sector is open to the public and to political processes. What is done in the public sector is public. In the private sector there has been a push towards such issues. A cooperative Health Maintenance Organization, located in Seattle has organized a public annual meeting consisting of members belonging to the organization. During one of their meetings resolution was passed that a study of comparable worth be done. Dr. Remick points out that over one third of the time was spent discussing the issue of comparable worth, thus showing its growing significance in the work environment.

Concerning unions, Dr. Remick underscores the fact that although unions may be dying in heavy industry,they are quite alive in the public sector and there is a significant amount of unionization in female dominated jobs. This is due to the fact that employees are asking for comparable worth studies.

Job evaluations are conducted for various reasons. Personnel departments have strong feelings surrounding job evaluation. Personnel departments hesitate in this area because job evaluations tend to upset the status quo. Dr. Remick advises that personnel departments make a better appraisal of the idea of job evaluation and also ask some questions concerning the conduction of the process. No real research concerning job evaluation has been done for the past twenty years. There should exist some clarification of the job evaluation system and the types of questions it raises and how it relates to salaries. The search for bias in the job situation should be a part of job evaluation and should be easily found.

The application os job evaluation systems has shifted. Traditionally the items chosen to be evaluateu included effort, skill, responsibility, and working conditions. A more thorough evaluation would include such factors as bias, interpersonal skills, factor relations, the application itself, and salary application. Dr. Remick used a few examples to clarify the items she listed under a more in-depth investigation. Factor relations would include such issues as assigning more weight to types of activities found in male dominated jobs such as heavy lifting. As for interpersonal skills it was found that the more interpersonal skills required, the less the salary was. Female dominated jobs usually involved interpersonal skills to a great extent. Also, the application of the whole system must be free of biased raters if it is to be effective at all. There are so many strong assumptions about the types of work men do and the types of work women do. The pictures of a job for a man and a job for a woman have already been painted for us. Women too are caught up in this present frame of mind concerning jobs. One example of this involves working conditions. A group of tree trimmers described their job as difficult, dangerous, and dirty and further classified it as a male job because of these characteristics. Dr. Remick stresses that the idea that males do the "dirty jobs" is a fallacy. Some workers function in a dirty environment and it is perfectly acceptable. This is true for auto mechanics. However, women too work in dirty environments but their work is centered around making the dirty or unkempt environment clean. An example would be nursing. Nurses usually describe their environment as being clean but actually nurses work at keeping their dirty environment a clean one. Dr. Remick questions how much of what goes on in personnel is based on the sex of the person doing the job.

The two main systems of job evaluation, to use as outlined by Dr. Remick, are non-priority and policy capturing. The non-priority system is mainly utilized by organizations which avoid statistical and mathematical procedures. It is a system in which a job evaluation idea is brought in by someone else and used if the organization feels it is a good one. Policy capturing incorporates devising one's own system based on statistics. That is, the organization finds its own factors and ratings and does its own study. The choice between the two systems is a political issue in the opinion of Dr. Remick. There are many issues to consider when choosing a system. Job evaluation is a difficult process. As an example, in the public sector there is a heavy emphasis on people and services. Assessing their value is a rather involved and intricate process. Much responsibility both for the job evaluators and the employees being evaluted is at hand. Such jobs as teaching, police work, and prison work are very sensitive to job analysis.

Dr. Remick gives a brief history of job evaluation and comparable worth beginning in 1973 when a study by Willis & Associates of the state of Washington was undertaken to look at management jobs. In 1974, the State Women's Commission asked for a study for various differences in sex segregated jobs. This was the first study in the country done using job evaluation to study sex segregated jobs. A total of one hundred twenty-one jobs was studied. The results were that the governor recommended that something be done about the issue. In 1976 another study was undertaken. The governor at the time left office and included seven million dollars in his proposed budget for the further development of this issue. The election of the new governor brought about the loss of that money. It was taken out of the fund. Thereafter, legislative bodies and other influential people declared the issue a very important and pressing one but never went beyond that point. Finally in 1982, the first semi-serious consideration of the legislation took place. The legislation went quite a way through the committees but finally died at the end. Two items which were passed however included a state commitment and implementation of comparable worth over the next ten years and that in July of 1984, all jobs twenty percent below the average salary line would earn a salary increase of one hundred dollars a year. Ten years and one hundred dollars a year are a lot of stretching according to Dr. Remick.

In the implementation of job evaluation and the issues of comparable worth there are many considerations to take into account. The salary to use in terms of range, where to start the studies, bias, errors, race, sex, etc. are among these considerations. Also to consider is the dire need for an evaluation of all jobs.

Dr. Remick suggests that the way to begin correcting the system is to move the salary line of women up to the level of men's. The unions' method of changing only average differences in salaries produces scattered results and does not correct the system. The unions uphold the idea that differences in salaries are based on supply and demand in the market. Overall there is said to be a thirty-one percent difference between what women are paid and what men are paid.

The costs of this process are astonishing. Bringing female dominated jobs and those jobs which are associated with them to equality will cost approximately sixty-six million dollars. It should be noted that this includes only the female dominated jobs.

Comparable worth is a hot issue for the future. Some firms are finally beginning to respond to this issue at least in presentation to win contracts. Comparable worth's real potency is yet to come.