DOCUMENT RESUME

ED 337 478 TM 017 292

TITLE	Proceedings of the 1981 IPMAAC Conference on Public Personnel Assessment (5th, Denver, Colorado, May 10-14, 1981).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	May 81
NOTE	74p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Assessment Centers (Personnel); Computer Assisted Testing; *Evaluation Methods; Job Analysis; *Job Performance; Management Information Systems; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; Predictive Measurement; *Test Use; Workshops
IDENTIFIERS	International Personnel Management Association

ABSTRACT

The International Personnel Management Association Assessment Council (IPMAAC) is a section of the International Personnel Management Association dedicated to the improvement of public personnel assessment in such fields as selection and performance evaluation. Before the IPMAAC's fifth annual conference in 1981, four workshops were conducted on the following topics: developing job-related minimum qualifications of training and experience; two computer systems for assessment use; and managing a test development unit. Author-generated summaries/outlines of papers presented at the IPMAAC's 1981 conference are provided. "Presidential Remarks" by J. P. Springer are reviewed. The keynote address is "Burden of Proof/Burden of Remedy" by M. Novick. The following paper sessions are summarized: "Alternative Selection Procedures"; "Setting Passing Points"; "Practical Procedures for Large and Small Agencies"; "Using Computers in Personnel Systems"; ""Measurement in Police Settings"; "Fine Tuning of Selection Instruments"; "Issues in Job Analysis"; "Test Validity and Utility"; "Personnel Adaptations of Video"; and "Considerations Concerning Discrimination." The following symposia are reviewed: "The Police Career Index, Tribulations and Trials in Des Moines"; "Executive Candidate Selection--Assessment Center Methods"; "Unassembled Examining--Current Approaches in the Federal Government"; "The Structured Interview--Dead or Alive?"; "The Development of Job-Related Medical Standards"; and "Survey of Basic Item Analysis". Three invited addresses include: "Contemporary Personnel Psychology--An Overview" by T. Hunt; "Truth-in-Testing Legislation" by R. Brown; and "Selection Research as Seen by a Personnel Director" by C. Grapentine. One other paper is summarized: "Assessor Training: The Leaderless Group Discussion". A microcomputer demonstration session is reviewed. (SLD)

IPMA Assessment Council

IPMA Assessment Council

Proceedings of the

1981 IPMAAC Conference

on

Public Personnel Assessment

May 10-14, 1981

Denver, Colorado

Presentation Index

INTRODUCTION

ABOUT IPMAAC

The International Personnel Management Association Assessment Council (IPMAAC) contains over 500 psychometric specialists, personnel psychologists, and personnel staffing specialists dedicated to the improvement of public personnel assessment in such fields as selection and performance evaluation. The Assessment Council evolved from a Selection Specialists Symposium sponsored in July of 1976. The one hundred and fifty individuals participating in that session supported the establishment of an organization to further intergovernmental communication and cooperation in the area of assessment, with the intention of improving selection practices. In April of 1977, the first IPMAAC annual conference was held in Kansas City, Missouri.

ABOUT THE CONFERENCE

The annual IPMAAC Conference is a major source of communication of ideas among assessment professionals in IPMA. The fifth annual conference, held in Denver, hosted over 100 presenters and approximately 180 attendees.

ABOUT THE PROCEEDINGS

The summaries of presentations in the Proceedings were prepared by members of the IPMAAC Education and Training Committee and others acknowledged below. The purposes of the summaries are to indicate topics addressed and to summarize important points.*

Some of the presentations are being prepared also for publication in their entirety as monographs or as articles.

Contributors

Bob Marshall
David Lookingbill
Kevin Love
Doris Maye
Nancy Abrams
Jan Klei
Charles Schultz
Barbara Showers
Bob Shoop
Phil Ferrara
Louis Laguardia

*NOTE: While every attempt has been made to accurately represent the presentations, persons wishing to quote results should consult directly with the original author(s).

WORKSHOP I

How to Develop Job-Related Minimum Qualifications and Ratings of Training and Experience

Leaders: Nancy Abrams,* U.S. Office of Personnel Management, New York Region
Louis M. Laguardia,* U.S. Office of Personnel Management, New York Region
Leroy Sheibley, Pennsylvania State Civil Service Commission

The purpose was to expose participants to several methods and theories of minimum qualification and T&E development and to give participants an opportunity to construct a T&E with guidance and critique from the workshop leaders.

The presentation on minimum qualifications began by identifying some of the traditional uses and abuses of minimum qualification statements. During the presentations, the following definitions of an MQ were offered:

1) An MQ is any specific characteristic or attribute which job applicants are required to have in order to be allowed to compete further in the selection process.

2) MQs should reasonably sort out those applicants who have a reasonable chance of performing or learning to perform the job from those who have little likelihood of success.

3) "Minimum qualifications required of job applicants should identify those elements of training, experience, special skills, or other personal attributes which are <u>essential</u> to performance of the job, and which must therefore be possessed by an applicant before he can reasonably be expected to perform a job satisfactorily." Tennessee State Department of Personnel. <u>Technical Standards for Determining Minimum Qualifications and Examination Weights</u>.

It was also recommended that a "good" MQ should possess the following characteristics:

Objectivity--no subjective standards;

Validity--clearly linked to specific work performed or specific job requirements;

Reliability--judgments should be consistent;

Acceptability--the logic and/or validity evidence should be in a form which can be understood by unions, operating agencies, applicants, etc.;

Have some cost/benefit value;

*Dr. Abrams is now a private consultant. Mr. Laguardia has joined the Port Authority of New York and New Jersey.

Be verifiable; and

Be developed in such a way that all reasonable options are considered.

It was also stressed that, like any other selection device, minimum qualifications need to be based on the results of a careful job analysis. The recommended method for developing minimum qualifications was to ask subject matter experts to specify examples of education and/or experience which would provide demonstration of possession of the required entry-level knowledges, skills and abilities.

After a short presentation on the background of training and experience (T&E) ratings, the following standards for T&E development were offered:

1) Based on job analysis;

2) Evaluate most important entry-level work behaviors or job requirements that differentiate superior workers from adequate workers;

3) Use supplemental form to collect information from applicants;

4) Use structured and well-defined rating procedure (e.g., behaviorally defined scales) for personnel selection specialist or SMEs to score supplements;

5) Be reliable, valid and have no adverse impact.

Working systems for T&E ratings in the states of Pennsylvania and Wisconsin were reviewed as were a self-rating task checklist and the B.R.E. Exam Preparation Manual. Both of the latter two systems are used to some extent in the Federal system.

In the afternoon, four groups were formed and a T&E rating guideline and questionnaire was developed for Carpenter by two groups and for Management Analyst by the other two groups. The afternoon concluded with a critique by the workshop leaders of each group's product.

WORKSHOP II

CODAP System 80

Leader : Doug Goodgame, Texas A&M University

This workshop acquainted personnel administrators and specialists with CODAP System 80. System 80 is an extensive series of computer routines designed to process data for job analysis, item banking and validation.

Data bases resulting from occupational surveys have three characteristics that make them somewhat unique: 1) The data bases tend to be very large. There may be several hundred to several thousand incumbent workers in a study with hundreds of observations recorded per worker. 2) The data base contains two distinct types of data, worker profile or background data and task responses. 3) Subsequent data processing and analysis tends to divide the original data base into numerous groups of data each requiring separate study. This situation places unacceptable demands upon commonly used statistical packages and forces analysts, who use such packages to process occupational survey data, to restrict the scope and bounds of planned analysis. The Air Force recognized these problems in the 1960s and initiated the development of the CODAP System to process occupational survey data. CODAP is an acronym for comprehensive data analysis programs.

System 80 is a redesign in integration of the large number of diverse programs for CODAP users which had developed over the years. It was developed to meet the following requirements:

1. The system should be flexible: new processing and display requirements should be easily developed by persons without special training in programming.
2. The program should be adaptive: the system should be able to process data other than relative time spent values without modification.
3. It should be easy to use: apprentice job analysts should be able to use the system to make routine runs without extensive training.
4. All data in the system should be accessible: any program or routine would be able to access any type of data in the data base.
5. The system should have high capacity: data storage limits should be expanded to meet current demand.
6. The system should be transportable: the system should be operable on any main frame equipment with minimum modification.

A data base management approach gave the job analysts the ability to assess any data, manipulate it in preparation for processing, process it using a wide variety of computing routines and display results in easy-to-read forms to analyze work. In this manner the job analyst can think of the data as raw material for building data summaries. To do this, the job

analyst uses English-like sentences to invoke the operation of certain procedures which pull data from specific locations and process it for display. The key to understanding this approach is: 1) a knowledge of the conceptual arrangement of data in the computer; 2) knowledge of CODAP language statements that invoke data processing and reporting; and 3) knowledge of sample formats for displaying results. These are the three knowledge requirements which a job analyst needs in order to process occupational survey data for analysis.

The areas covered in the workshop were: data collection procedures--the use of task inventories, a comparison of utility of various job analytic methods, basic operating characteristics of CODAP System 80, and position administration including selection, training, job evaluation and classification.

WORKSHOP III

PAQ System II--Emphasizing Personnel Selection Applications

Leader: Robert C. Mecham, Ph.D.

A. Job Analysis as the foundation for personnel administration decisions.

B. Differences between various job analysis procedures and products were discussed in terms of:

1. Worker-oriented vs. job-oriented methods
2. Numerical vs. written products
3. Structured vs. unstructured procedures
4. Standardized vs. customized procedures

C. The Position Analysis Questionnaire (PAQ) was introduced as a standardized, structured, numerically rated, and worker-oriented type of job analysis. Types of job elements, rating scales, and job profiles were illustrated.

D. Procedures for collecting and checking the reliability of PAQ data were described including the use of:

1. Inter-rater reliability analysis
2. A modified paired-comparison analysis and
3. Data base referencing

E. Computer processing options were described which included:

1. Derivation of job profiles
2. Statistical comparison of job profiles for two or more jobs to determine degree of similarity
3. Derivation of job families using cluster analysis
4. The prediction of aptitude test means, validity coefficients and cutting scores for various jobs using job profiles as predictors
5. The prediction of job evaluation values for jobs from job profiles
6. The process of searching for a job data to determine various career paths for employees
7. The vocational guidance procedures using a combination of PAQ data for jobs, vocational preference data from the individual gathered using the Job Activity Preference Questionnaire (JAPQ) and aptitude scores.

F. A computer-augmented personnel system currently under development which uses PAQ data was described. The system is expected to be capable of the following:

1. Collection and recording of application blank, vocational preference and aptitude data from job applicants using a computer terminal
2. Validation of predictors against performance appraisal, tenure and other performance criteria
3. Prediction of jobs an applicant is expected to perform best

4. Optimizing placement of a number of applicants on a number of jobs
5. Identification of discrepancies between employee validations and job requirements for use by employment interviewers
6. Career path searches
7. Vocational planning, etc.

WORKSHOP IV

Managing a Test Development Unit

Leaders: Chuck Schultz, State of Washington
Charley Sproule, Pennsylvania State Civil Service Commission
Ron Ash, University of South Florida

Workshop presenters led participants through consideration of several major management problems of a test development program. For each topic, presenters highlighted their experiences, gave examples of working systems in use, and participants shared and compared their ideas and experiences with each other. The result was a useful opportunity to learn new approaches and evaluate one's own techniques in the many responsibilities associated with managing a test development unit.

Topics covered were:

- Defining and measuring productivity and implementing procedures to enhance productivity.
- Staffing for test developing, including recruiting and selection criteria.
- Planning and organizing--setting priorities, scheduling projects, and coordination with other units and agencies.
- Techniques for small jurisdictions.
- Training test development specialists and nonspecialists.

Chuck Schultz began by discussing productivity and methods of measuring and encouraging it which he has implemented in his unit. He emphasized the need to establish and maintain priorities and the need to track progress toward goals through reporting systems. Discussions revealed wide differences in expected lengths of time to complete test development projects, from days to months for similar projects. Discussion also dwelt on maintaining standards of quality and establishing criteria for test quality. Job relatedness, correct grammar, lack of bias, and readability were identified and discussed as essential criteria. Recruiting of staff for test developing was also discussed, with emphasis on selection of criteria and some debate on need or desirability of college degree or higher education.

Charley Sproule presented the large scale approach for exam scheduling and test development priorities used by the State of Pennsylvania. Pennsylvania projects a schedule for the entire year through an annual survey of agencies and updates it quarterly. It categorizes projects in advance by levels of commitment of staff time, from continue "as is" through six months criterion-related validity studies. It also employs a category of exams called "accelerated examinations" which can be developed

outside the normal process for small classes or those where few appointments are expected. He presented an employment data system which is used to track agency needs, average work force availability, candidate availability through tests, and appointments for all job classes. He also discussed staff training needs and the programs and resources available through his own agency and through regional consortia.

Ron Ash discussed coordination of work within the organization and especially the effectiveness of various job analysis methods for multiple organizational purposes. He presented the results of a study he and Ed Levine conducted which was a survey of experienced job analysis users concerning effectiveness and practicality of seven common methods, such as job element, functional job analysis, CODAP, PAQ for eleven purposes, such as job description, classification, selection, performance appraisal. They found significant differences between the methods in both effectiveness and practicality. The results were expected to be published soon.

Ash also led discussion of techniques for small jurisdictions, such as purchasing tests, structured interviews, and training and experience evaluations. Programmable calculators were discussed and compared as economical means of data analysis. He provided descriptive resource materials from test publishers and calculator manufacturers.

Finally, the topic of discussion turned to implications of validity generalization and synthetic validity as other possible alternatives to reduce workload and increase test use and flexibility. From considerations of productivity and work scheduling through validation techniques, this workshop presented a great deal of useful information, examples, and shared experiences to assist test unit managers in evaluating and improving their own approaches.

PRESIDENTIAL REMARKS

James P. Springer

Because of a promotion to a broader personnel management position during the year, Mr. Springer was able to share a new perspective on the role of merit selection in personnel and in government. He emphasized the lack of priority of selection from the viewpoint of elected officials and the lack of understanding of the role which quality selection plays in cost corscious and effective government. He emphasized the critical need for selection professionals to take the time to translate validation research into dollar benefit and demonstrate that better tests yield better employees.

Some specific ideas he proposed were:

- doing more cost effectiveness research and reporting it in publications such as Public Administration Times, Public Personnel Management, rather than in the Journal of Applied Psychology;
- using the concept of life cycle costing as the purchasing field does to describe utility of personnel selection decisions in terms of total life cycle costs.

He cited a Milwaukee study which showed that quality selection was more cost effective. The study assessed the cost of discipline and absenteeism between high and low scoring individuals and found a savings of over $100,000.00 per year in 217 hires. A follow-up study showed that savings were still being made after three years.

Mr. Springer observed that we have not pursued this line of communication with our own agencies often enough, but that with increasingly tight budgets, it will be essential for us to talk in dollar terms to continue the progress of personnel selection in the coming years.

"What I have tried to do this morning is to raise questions about where selection is on the list of priorities as they are retranslated in terms of limited financial resources. I have tried to reemphasize the values underlying selection and to point out that these values are not readily apparent--that we must designate some of our energies to competing for limited resources.

"With the demise of IPA and CETA, we should be convinced that less money will be available. It is too bad that we do not have IPA to continue the efforts in the manner in which we have been involved in the recent past. We have to recognize the value of IPA and the federal officials who helped and contributed to our efforts. But our efforts at cooperation will continue, I am sure.

"If you think of it, this group represented here is the nucleus of the public personnel selection area, now and for the immediate future. The thoughts you form and the actions you take will determine the progress of

public personnel selection in the coming years. I hope that my brief remarks have raised some questions in your mind and a predisposition to act at this critical time.

"The idea of 'let's talk in dollars' does not emphasize the other many positive values of an effective selection program, but for the time being we have to speak the language of the realm."

PAPER SESSIONS

ALTERNATIVE SELECTION PROCEDURES

Chair: Kaye Evleth, Los Angeles

Discussant: Cindy Cook, Indiana

The Search for Alternative Selection Procedures

Terry McKinney, City of Phoenix

Fifty-three jurisdictions responded to a survey about the attempt to conform with the Guidelines' cosmic search requirement. Most of the agencies do not conduct a thorough search for alternatives although they understand that it is required. Most of them indicated that they did not have time to make the search. Some felt that since they use content validation strategy rather than criterion-related strategy, which yields apparent indicators of the magnitude of validity, they were exempt from the search requirement.

When a search is made, agencies will most often canvass other public jurisdictions and are unlikely to go to private companies or to universities.

The author noted a confusion among authorities on issues related to the use of tests that show adverse effect. Courts, guidelines, Merit System Standards, and professionals seem to be pursuing different objectives with little regard for the pronouncements of one another.

Experience Requirements in Selecting Employees

Richard D. Arvey and Evelyn E. Miller, University of Houston

A literature review shows that most experience requirements, usually global requirements, are not correlated with job performance. In the few instances in which job requirements are correlated with performance, the experiences are specific, quantifiable and job-related.

Courts tend to uphold experience requirements, although decisions are not consistent. Experience requirements tend to be upheld in cases where there is great human or economic risk and higher level jobs. They are less likely to be upheld when there is adverse effect on protected groups.

Development of Pre-Employment Questionnaires

David C. Myers and Sidney A. Fine, Advanced Research Resources Organization, Washington, D.C.

Self report questionnaires were developed to assess experiences that were related to job analysis results from several job families. Items were edited to eliminate bias. A scoring system was devised to discriminate among jobs within families in terms of degrees of relevance.

Self-Assessments of Mental Abilities

Michele E. Fraser and Richard D. Olson, Personnel Decisions, Inc., Minneapolis

Single-method, single-sample, single-criterion studies of self-assessments produced inconsistencies which were hard to explain. These authors sought to answer the question of when to use, rather than whether to use, self-assessments. Two multi-trait, multi-method studies asked men and women to rate their own mental abilities.

Two methods were used to obtain self-assessments. For "skill ratings," participants indicated how easily they could perform each of a number of activities related to the mental ability. For "ability ratings," participants indicated how they compare with a normative group. In Study 1, the ability ratings had higher validities; in Study 2, the skill ratings had higher validities.

Convergent validities were only about .40 on the average, ranging from .17 to .73. Discriminant coefficients were generally lower by a reasonable amount.

Sex differences were variable and sometimes large. For example, women scored significantly higher on the clerical ability test and the skill ratings showed women better to an appropriate degree, while the ability ratings implied that the men had slightly more of this ability. The opposite pattern of results appeared for spatial ability.

There was a general tendency for people to be lenient in rating their own abilities. However, this can be controlled by standardizing the ratings. The results suggest that we should continue to look at men's and women's self-ratings separately.

SETTING PASSING POINTS

Chair: George Nelson, Denver

Criterion-Referenced Standards Based on Theory and Experience

Maureen Kaley and Sandra Singer, Professional Examination Service

The presenters focused on recent developments and findings pertaining to criterion-referenced methodologies. A review of the passing point setting rationale and methods was presented, with discussions of recommended guidelines and legal requirements for setting passing points. An annotated bibliography of 50 references was made available to participants.

The Nedelsky, Angoff, and Ebel methodologies were described. The Nedelsky and Angoff were based on the judges item-by-item estimates regarding the ability of minimally competent examinees to either eliminate distractors (Nedelsky) or get the item correct (Angoff). The variance of the judges' estimates of total score was assumed to approximate the variance which minimally competent examinees received. The Ebel method was based on classification of items into importance by difficulty categories. The judges estimated the percent in each category which a minimally competent examinee could be expected to get right. The method resulted in a single passing score.

An account of the experience of a group of environmental health specialists (radiation protection technologist) who made the transition from using a norm-referenced passing point determination to implementing a criterion-referenced approach to reaching a passing point was described. Considerable emphasis was placed on group discussion of the job requirements and giving of rating feedback. The group was not required to come to consensus on ratings, but all divergent ratings were brought to their attention and discussed. The group also considered standard error of measurement and adjusted their final recommendation down 1 SEM to allow for measurement error.

Evaluation of Panel Review Method

Barbara Showers, Wisconsin Department of Regulation and Licensing

Data from three item reviews and test administrations of the Wisconsin Real Estate examinations were analyzed to determine inter-rater reliability, stability of results over time, validity of judgments, and public credibility of the panel review procedure for setting passing points. Some findings of the study were:

1. The combined judgments of the groups of four to six raters which were used to establish passing points were typically found to be reliable, with an average reliability of .73 over 12 rater groups. There were no apparent differences in reliability of ratings for national and state items, or items with and without answer key provided.
2. All three recommended passing points for each of the four tests were within about five percentage points, with one exception on one test.
3. The implemented passing point in the exception above was determined by the policy of overriding rater judgment by actual rater performance when rater performance was lower. The policy resulted in maintaining the stability of the passing point over time.
4. There were significant differences in average ratings between raters in every group. The differences tended to average out in most cases, but their presence emphasized the importance of representative sampling and the possible need to refine the use of the scale of probability values.

Documentation of Ranking and Minimum Cutoff Score Use for Content Valid Tests

William Howeth, McCann Associates, Inc., Huntingdon Valley, Pennsylvania

Cutoff/ranking (C/R) analysis is a procedure for quantifying the judgments of subject matter experts to document the appropriateness of using written test scores both to rank candidates and to establish a cutoff score. C/R analysis requires a population of incumbents performing the job for which the test is to be used, as well as a population of lower level employees who are eligible to compete for promotion to that job. In addition, there should be available higher level superiors who qualify as SMEs for both jobs because they are intimately involved in supervising and directing the work of both the employees who are eligible to compete for promotion as well as the incumbents performing the job for which the test has been developed.

The procedure involves first asking the SMEs to make judgments as to how eligible candidates would be distributed in terms of job performance, if all of the eligible candidates were promoted. (Four carefully defined performance levels are used.) Next, the SMEs are asked to actually review the test questions used in the written test and make the following judgments:

1. Is the knowledge or ability measured by the question related to successful performance of the job?
2. Four separate judgments on the likelihood that performers at each of the four levels would possess the knowledge or ability; i.e., be able to answer the question correctly.

Each SME rates the likelihood of individuals at each performance level answering the question correctly using a three-value scale. The ratings are then translated into item difficulties. Predicted mean test scores for each performance level are then generated and compared to each other. The extent to which higher test scores are predicted for higher performance levels supports the test's use for ranking.

Using assumptions of normality, predicted score distributions are developed for each performance level using both the predicted test score data and the predicted percentage of eligible candidates falling into each performance level. The range of written test scores where the score distribution for the acceptable performance level overlaps the predicted score distribution for the unacceptable performance level is identified. Possible minimum cutoff scores are then discussed in terms of the risk of passing potentially unacceptable performers versus the risk of eliminating possible acceptable performers.

Aspects of Inter-Judge Variability in Setting Criterion-Referenced Passing Points on a Credentialing Examination

Leon I. Smith, Professional Examination Service

Seven members of an examination committee in a professional field implemented the Angoff procedure after constructing a form of a licensing examination. Following an open-ended discussion concerning an operational definition of the "minimally competent" professional, the members of the committee rated ten questions from the item bank as a sample exercise. The judges were then polled and discussed discrepancies and the reasons for their probability estimates. The members of the committee then rated each of the items on the newly constructed form. After completing their ratings, but before receiving feedback of results, the judges were asked to respond to rating scales concerning: a) their degree of comfort in recommending the passing point that would be produced; b) the need for additional groups and standard setting; and c) their ability to implement the operational definition of entry-level minimal competency.

While the findings indicated that the committee felt comfortable with the procedure, the consensus was that additional groups should be involved in the judgmental process. Practitioners and licensing board representatives were the most frequently mentioned groups that should be included in the standard setting procedures. Perhaps of most importance, the degree of comfort in recommending the passing point that would be produced was related both to their judged ability to implement the definition of minimally competent and the actual variability of their item ratings. There was also some evidence that specific content expertise affects item judgment. Judges' ratings of items within their areas of expertise appear somewhat higher (implying a more rigorous standard) than their ratings from other parts of the test blueprint. Implications of these findings were discussed in terms of the need for and of developing a precise definition of minimal competency to reduce judgment error and the importance of selecting judges on the basis of matching content expertise to test specifications.

KEYNOTE ADDRESS

"Burden of Proof/Burden of Remedy"*

Melvin Novick
Professor of Education and Statistics
University of Iowa

What is discrimination in test use and who is going to do something about it? Many people react to the suggestion of discrimination without having a very good idea about what it is or how their reactions will affect the problem. Various definitions of discrimination have little in common, therefore, the generic term discrimination should not be used.

Statistical definitions of test bias are inadequate. Classical statistics, employing the concept of random sampling, are inappropriate for use in practical selection settings. An unquestioning application of statistics to test bias can lead to irrelevant conclusions. Different conclusions can be drawn from analyses of different combinations of the variables. Statistics should be employed with understanding. The understanding must come from a sophisticated and judgmental consideration of the problem.

The Uniform Guidelines are no help. They contain no real definition of the problem. No useful methodology is provided. The cosmic search burden is not enforceable.

Court decisions are not consistently related to congressional intent or to scientific knowledge. Personnel administrators make uninformed decisions and often quit using tests in favor of other selection devices that have less utility and greater adverse effect. It is not reasonable to ask each individual jurisdiction to satisfy the fleeting requirements of courts and compliance agencies. IPMAAC members have a vested interest in being more influential in the decisions made about testing. Test users should be mindful of factors that affect test performance and be mindful of the use that is made of test results.

*Published in Public Personnel Management, 1981, 10 (3), 333-342.

PAPER SESSIONS

PRACTICAL PROCEDURES FOR LARGE AND SMALL AGENCIES

Chair: Janet McGuire, Arlington County, Virginia

Discussant: Ann Stillman, Office of Personnel Management, Dallas, Texas

How to Develop Community, Interdepartmental and Assessment Division Cooperation

Fay Walther, Fort Worth, Texas

In a small jurisdiction, the Personnel Assessment Division, community leaders, and hiring departments must work closely together to facilitate the selection of competent city employees. This teamwork approach to selection encompasses the research expertise of the Assessment Division and the content knowledge of the community leaders and hiring departments. The approach systematically develops the criteria for job performance, weights the criteria by importance and frequency, develops anchors of job performance for the rating scales, and establishes inter-rater reliability.

The steps to implement this cooperative approach include public relations skills and sensitivity toward the hiring departments' staffs. The Assessment Division has a major responsibility to assist the hiring departments in meeting government guidelines by providing technical selection procedures. Therefore, an essential step is to create an image of assisting the departments rather than as a "hurdle to be overcome." This credibility facilitates the creation of mutual, job-related goals and selection procedures.

In conjunction with the Assessment Division and the hiring departments, community leaders with content knowledge of specific positions are requested to participate in the selection process. This utilization of community resources helps to overcome the problem of a limited number of professional personnel staff and enhances the accurate and comprehensive measurement of job criteria. The method was applied with the high level, specialized position of library director. The positive response develops the support of the community for the successful job performance of the new director.

The advantage of developing interdepartmental and community cooperation is increased efficiency. The process is a structured approach to develop job criteria, measurement scales, and the systematic training of raters. It is an attempt to achieve the "best of both worlds" in a small jurisdiction: research expertise combined with the practical job knowledge and support of the hiring departments and the community.

Job Elements for Task Clusters

Ron Ash, University of South Florida, Tampa

Ash identified the different foci of job analysis for different validation strategies according to the Uniform Guidelines. They were:

Content Validation: observed work behaviors, tasks, observed work products

Construct Validation: work behaviors, underlying constructs

Criterion - Related Validation: measures of work behaviors or performance representing important job duties, work behaviors, or work outcomes.

He then classified major methods of job analysis according to type of information obtained:

Task Based: functional job analysis, task inventory/comprehensive occupational data analysis programs, Department of Labor task analysis

Behavior Based: critical incident technique, position analysis questionnaire

Attribute Based: functional job analysis, division analysis questionnaire, job element method, ability requirements scales.

He went on to describe the application of a multi-method job analysis approach to a job analysis study of condominium managers in Florida. The methods used were task inventory, job element method (JEM), and position analysis questionnaire (PAQ). He found that the combined results were useful for several functions. By linking the JEM study to task inventory results, KSAs were identified and linked to job tasks, selection needs, training, and performance appraisal uses. Illustrations of resulting products were presented, such as the KSAs suitable for potential inclusion in condominium management training programs by job task dimensions.

Selection for One-of-a-Kind Job Classes

Carla Swander, Metro, Seattle

A large part of personnel selection applies to one or two position classes. Ms. Swander deals with the issue of whether or not to do a job analysis for these classes, since she points out that most common methods, including interjurisdictional studies, do not work for these classes. In her position with Metro, she found that selection specialists need to teach generalists how to get job analysis information. She has developed a training manual for "informal job analysis" to communicate with the generalists and teach them how to do interviews for selection purposes.

The manual includes such things as: who to interview, how much time it will take, how to take notes, how to help the subject matter expert determine the importance of job elements, recognizing credentialism when they hear it, focusing on first-day needs of the job and avoiding trained skills, helping the conversation to stay on track, helping others to organize their thoughts, finding the real meaning of "you have to have a college education," spotting redundancy, general lines of questioning, and the pros and cons of particular selection methods. The manual emphasizes that they have the responsibility to define the major elements of the job and to work with the subject matter expert to define subelements of the job.

In her experience, she has found that she must actually work with the generalists to show them how it is done, but that this has been a very satisfactory approach to the problem of identifying critical job content and developing pragmatic selection tools for one or two position job classes.

USING COMPUTERS IN PERSONNEL SYSTEMS

Chair: Ed Cole, Sacramento Municipal Utility District, California

Discussant: Bob Shoop, Missouri Personnel Division

Job Evaluation Through Computer Usage

Nicholas F. Horney and Marvin G. Dertien, Salt River Project, Phoenix

Job evaluation has recently received a great deal of attention from those who allege that current systems of evaluating job worth contribute significantly to systematic wage discrimination. Further attention has been given to job evaluation in light of the Equal Employment Opportunity Commission's emphasis on the theory of comparable worth.

The EEOC commissioned the National Academy of Sciences (NAS) to study the feasibility and desirability of developing job evaluation methods that are fair and objective. A preliminary report from NAS pointed out that most published research on job evaluation methods is approximately thirty years old. Most of this past research dealt with traditional job evaluation methods as point systems, factor comparison systems, abbreviated versions of both point and factor comparison systems, and reliabilities and factor structures associated with jobs.

The research on a new job evaluation technique, reported in this paper, deals with work carried out over the past several years. However, with the ever-expanding technological advances in the computer field, much of the work reported was conducted within the last year.

The research concerns the development of a job evaluation technique which is nondiscriminatory, objective, and closely linked to computer technology. The technique, which is referred to as FACTS (Factor Analysis and Calculation

Technique for Supervisors), utilizes self evaluation by employees, yields a multiple correlation of .9741 with wage rates, produces understandable results, and is closely linked with the computer for rapid results. In addition to providing quick job evaluation results, the computer is utilized for analyzing the data with nonlinear multiple regression programs provided in UCLA's statistical package, BMDP (Biomedical Computer Programs, 1979 version). The result is a high multiple correlation coefficient, a relative low standard error, and no negative weights to try to explain to employees.

Part of the objective for the FACTS program was that of communicating the results to the employees. Previous job evaluation approaches utilizing multiple regression weighting of factors include the Position Analysis Questionnaire (Mecham and McCormick, 1969) and the Position Description Questionnaire (Gomez-Mejia, Page, and Tornow, 1979). Both of these approaches apparently used stepwise multiple regression for data analysis. However, stepwise regression was rejected in our research because of two perceived disadvantages: (1) stepwise regression tends to preclude redundant variables from entering an equation, which can result in an over emphasis of a single factor; and (2) stepwise regression can yield a negative weight for a factor whose simple correlation is positive. To solve this problem, a special regression function was written into a subroutine which was linked to the BMDP package. This modification allowed us to develop an equation with all positive weights. The equation yielded a multiple correlation of .9741 and a standard error of .73 salary grades. By comparison, an equation consisting of both positive and negative weights yielded a multiple correlation of .9796 and a standard error of .64 salary grades. While all positive weights slightly increased the error in the equation, it was considered an acceptable tradeoff in order to eliminate a major barrier to explaining the job evaluation system to employees (i.e., the negative weights of certain factors).

Although much of this work could not have been done as rapidly without computer facilities, this author believes that advances in the computer field are making applications available for human resources research work previously thought to be too time consuming or too costly to be feasible. Therefore, further research is being conducted in our organization concerning the linkage of the computer and human resources research.

Computerization of a Test Development System from the Perspective of a Testing Agency

Maureen Kaley and Sandra Singer, Professional Examination Service

With rising inflationary costs impacting every aspect of the economy, many agencies involved in competency assessment are taking a hard look at where and how to cut costs. A major problem confronting those in the public sector who have such responsibility is how to reduce expenditures without impairing the quality of service provided to the public.

In response to the need to "at the very minimum" hold down costs, one testing agency made the decision to computerize its services. This paper consisted of a detailed description of how the computerization was implemented.

The paper had three main sections. In the first section, the factors that influenced the decision to computerize were presented. The advantages and disadvantages associated with each factor were discussed, e.g., the impact on staff productivity, the potential for morale problems, and cost effectiveness in the immediate and distant future.

In the second section of the paper, the new computer products, including hardware equipment and software, were discussed with emphasis on how they were used to computerize the item bank and the test development and scoring procedures.

The third section consisted of a discussion and analysis of the transitionary period; that is, the time period during which existing test development systems became computerized.

There was a discussion of how various divisions within the agency were prepared for and participated in the transition and how one examination program fared during the transitionary period.

Operating Characteristics of CODAP System 80

Doug Goodgame, Texas A&M University

Personnel analysts are entering an age where computers will be used more extensively in personnel work. One of the problems confronting the use of computers in personnel administration is the personnel analysts' lack of training in using the computer to process and manipulate data. (Displaying information from personnel records on a display device is not a data processing function.) Processing data is normally dependent upon availability of statistical packages with routines assigned to specific functions or programmers who use a language such as FORTRAN or COBOL to create special routines for processing data. In either case, personnel analysts are confronted with difficulties that prevent effective use of the computer. In the first case, the number of variables in personnel work such as job analysis is too large for conventional statistical packages. Secondly, personnel analysts are not trained to function as programmers and often have difficulty communicating with data processing personnel to obtain needed programming services.

An effort is presently being concluded which will help overcome these and other difficulties which personnel analysts encounter in using the computer. The Occupational Research Division at Texas A&M University is under contract with the Navy Occupational Data Analysis Center (NODAC) to create a database management system that will utilize an interpreter to invoke routines for processing occupational personnel data. This is the result of a major top-down redesign of the CODAP System developed by the Air Force. (CODAP

is an acronym for Comprehensive Occupational Data Analysis Programs and consists of a set of programs designed to process job analytic data. (Refer to the symposium entitled "Computer-Based Job Analysis--Some Innovative Applications to Personnel Management" at the 1980 IPMAAC meeting.)

CODAP System 80 is the result of this effort and will operate in such a manner that persons untrained in computer use will be able to learn to perform complex data processing operations. A language of about forty English-like words and a few symbols, when used to create statements resembling sentences, will then be processed by a complex interpreter in CODAP System 80 to invoke routines to process data for analysis and interpretation. The primary requirement for learning to operate this system is knowledge of how to use System 80 words and symbols to create sentence-like statements. The presentation reviewed the background that led to development of CODAP System 80, its operating characteristics, sample data summaries used in job analysis, and an example System 80 statement to illustrate how a personnel analyst will process data for reporting and analysis.

SYMPOSIUM

The Police Career Index, Tribulations and Trials in Des Moines

Moderator: Michele Fraser, Personnel Decisions, Inc., Minneapolis

Presenters: Marvin Dunnette, Personnel Decisions, Inc.
Pierre Meyer, Personnel Decisions, Inc.
Reg Sipfle, Des Moines
Richard Keenan, Thompson, Nielsen, Klanerkamp & James

The symposium dealing with the Police Career Index (PCI) looked at its development and at the suit brought against its use in Des Moines, Iowa. The PCI is a paper and pencil test to be used in the selection and promotion of Police Officers, Sergeants, and Intermediate Commanders. It was developed in a national validation study by Personnel Decisions, Inc., of Minneapolis, funded by LEAA, in 1974-76.

The beginning work on the PCI was done in 29 cities of various sizes throughout the U.S. in which police administrators and personnel officials were interviewed in order to learn about the selection and promotion practices currently being used. Next, a series of workshops with representatives from four levels of police personnel (patrol officers, detectives, sergeants, and intermediate commanders) was conducted to gather critical incident information--approximately 2,500-3,000 incident reports were recorded. From this information Behavior Anchored Performance Description Scales for each of the four positions were developed.

In the next phase, an experimental battery of tests and inventories was formulated. These included background information, specially developed situational judgment questions, cognitive tests (such as vocabulary knowledge, deductive reasoning, and perceptual speed and accuracy), opinion and self-description statements (such as items from the CPI and MMPI), and preferences (such as school subjects and hobbies).

The results of this battery were obtained from police officers in nine cities along with supervisory ratings. Various types of analyses were conducted on these data, such as: an evaluation of the dimensionality of the criterion information; a comparison of multiple supervisor ratings; and factor analysis with various rotations. Empirical scoring keys were developed from item analysis (using criterion and modal response weightings) and cross validation (using a Monte Carlo approach to get holdback validity coefficients and also a random criterion strategy). Thus, a series of separate keys for each of the four areas was obtained.

In the early 1970s, the Des Moines Civil Service Commission was under pressure to abandon all written exams. Only police, fire and clerical positions had written exams, but even in these cases the results received little weight in comparison to oral exams. However, in 1972 the city was

was charged with sex discrimination because of the weight and height requirements for police officers. The results of the charges were that all police exams had to be validated.

At this point, the Civil Service Commission found out about the work being done on the PCI. In 1978 the PCI was given as part of the promotion examination for police sergeant in Des Moines. However, two police officers who failed the exam challenged its use before the Civil Service Commission, the State Civil Rights Commission, and the District Court. They charged that the exam was discriminatory, had adverse impact, did not meet Federal Uniform Guidelines on Employee Selection Procedures, and they should be allowed to examine all test material including the answer key. They sought a permanent injunction against the use of the test results or any future use of the PCI.

Because of the seriousness of the matter, Personnel Decisions, Inc., decided to take an active interest in the case and was granted status as intervenor. Before the actual trial began, the issues became somewhat changed. The plaintiffs agreed that the exam was not discriminatory, there was no adverse impact, and the validity of the exam did not have to be proven.

A major complaint brought out during the trial concerned the appropriateness of some of the questions. The two police officers maintained that the questions dealing with biographical and psychological data were not job related. They said that, because this data cannot be changed, it is impossible to improve one's score and ever pass the test. They held that the cutoff score of t=50 (the mean score of the normative sample) was arbitrary and unfair They also maintained that the applicants had the right to examine all test material including the answer key.

In 1979, the Court ruled in favor of the Plaintiffs and a permanent injunction was granted. However, the City of Des Moines and Personnel Decisions, Inc., appealed to the Iowa Supreme Court, which in 1980 reversed the decision of the District Court. The reversal stated that the PCI did not violate Civil Service Commission rules, the cutoff point was established appropriately, and it was sufficient that applicants be able to inspect their answer sheets, but not the answer key, in order to preserve the security of the exam.

Since the time of the trial, revisions have been made on the PCI. These include removal of gender specific wording, removal of some of the items that had been found objectionable, discontinued scoring services for the Detective PCI, and development of a shorter, simplified report form.

Finally, in the discussion of the PCI and the Des Moines trial, the need was stressed to pursue vigorously legal challenges when they occur. The legal status of the consultants must be established, the services of knowledgeable attorneys must be acquired, and careful preparation for courtroom presentations must be made.

SYMPOSIUM

Executive Candidate Selection - Assessment Center Methods

Moderator: Lawrence S. Buck, U.S. Department of Agriculture

Presenters: Gary Brumback, U.S. Department of Health and Human Services
Warren Johnson, Jr., U.S. Department of Agriculture
Sandra Pilch, U.S. Department of Agriculture
Jack Clancy, Sacramento, California

Lessons Learned in Selecting Candidates for Executive Development

Gary Brumback

As head psychologist with the U.S. Department of Health and Human Services, Brumback presented an overview of the assessment center used for the department's Executive Candidate Development Program (XCDP). He also expressed his concerns regarding assessment centers and discussed some considerations for future change.

The primary objective of XCDP is to prepare selected staff members for senior executive positions. Successful participants in the two-year preparatory program are eligible for non-competitive placement in high-level managerial positions throughout the department.

The programs's selection process consisted of six hurdles, the fourth being the assessment center itself. Prior to the center, candidates underwent an elibility screen, a paper screen which included a review and evaluation by a pair of raters of candidates' applications and supplemental appraisal forms, and a three-part hurdle utilizing an interview, a group discussion and another paper screen. Following the assessment center, the remaining candidates underwent reviews by agency boards and the department's Secretary.

The assessment center employed four exercises--an in-basket, a group discussion, a problem analysis, and a leadership exercise. The private consulting firm which developed and administered the exercises consolidated 23 skill dimensions recommended earlier by another consultant with two department-suggested skills into 11 basic dimensions. A three and one-half day assessor training course stressing the observation of those skills preceded the actual assessment center.

Of the 236 applicants and nominees for XCDP, 60 passed the first three hurdles and participated in the assessment center. All of these individuals were retained for further review; subsequently, 42 candidates were selected by the Secretary for appointment to the developmental program.

Looking back, Brumback made several critical observations. He questioned the concept of XCDP--that is, do such contrived training environments

lessen the judgment and decision-making skills of candidates (who, he suggests, may be better served by experiencing more realistic job assignments)?

Brumback also cited what he felt was a weak job analysis in this case and emphasized the need for more job-specific dimensions. His own approach to the future evaluation of managerial jobs would be task-based employing the critical incident technique. Assessment exercises derived from such an analysis would be content validated using a quantitative estimate such as Lawshe's content validity ratio. Hurdles prior to the assessment center would be eliminated and all candidates would participate in a streamlined center.

Such an assessment center would stress the objective measurement of behaviors by using a minimum of exercises and dimensions to the fullest extent. Overlap in skills measurement would be reduced through fewer exercises and more precise performance dimension definitions; however, the observation of skills would be increased by requiring more than one assessor per candidate per exercise.

In summary, Brumback felt that the parts of the assessment center outshone the whole. Assessor training, the objective observation of skills in simulations, and detailed job analysis which affect all components of assessment centers can be utilized in a more cost effective manner as independent aspects of management development programs.

The Customized Assessment Process Developmental Needs Analysis

Warren Johnson, Jr. (presented by Lawrence Buck)

The Department of Agriculture's response to the creation of the Senior Executive Service in the federal government was the establishment of the Candidate Development Program (CDP). An integral part of this program was the Customized Assessment Process (CAP) which Johnson summarized in his paper.

CAP provides the means to assess individual training needs through extensive utilization of the assessment center technique. The target position in this process is a departmental executive functioning as a generalist.

Candidates apply for nomination to their respective agencies in the Agriculture Department. Agency nominations are reviewed by Program Evaluation Research Boards who select the participants for the assessment centers. The results of the center are then reviewed by an Executive Service Research Board which passes on entrance into the CDP.

The program, in addition to assessing managerial competence and developmental needs, supports the objective job analysis, orients participants to the department and its functions, and seeks to involve as many in-house people as possible.

Four simulated situations in CAP were specifically designed to replicate the critical elements of a generalist position. Nine factors, identified by a content validity study, are rated based upon candidate performance in the exercises. Six teams of three--a USDA training officer, an agency Senior Executive, and an independent psychologist-consultant--observe and interview the candidates.

The methodology includes a sorting, prioritizing, organizing (in-basket) test (S.P.O.T.), a problem analysis solution test (P.A.S.T.), a leaderless group discussion, and a one-on-one interview. At the conclusion of the entire process, each candidate, consultant, evaluator, and reviewer evaluates the process.

Johnson feels that the extensive use of department personnel is a morale builder and helps the process maintain a high degree of validity. He estimates that the utilization of agency, rather than outside, resources has saved $155,000 for the department in evaluating over 200 candidates.

Candidate and Assessment Profiles for the U.S. Department of Agriculture SES Candidate Development Program

Sandra Pilch

The Senior Executive Service Candidate Development Program (SESCDP) in the Department of Agriculture was begun in November 1979 with the nomination of 127 GS-15 managers to compete for thirty program vacancies. In describing the outcomes of the program's Customized Assessment Process (CAP), Pilch reviewed the biographical and assessment information for the eighty individuals whose executive skills were evaluated in the CAP.

She reported that most participants were male, middle-aged ($\bar{X}$ = 46.3 years) and mid-career ($\bar{X}$ = 20 years of professional work experience); all possessed at least a bachelor's degree. The average candidate held positions in two different occupational series, general administration and biological science.

The composite candidate received a "satisfactory" overall rating on his most recent performance appraisal and at least one USDA award for "outstanding" job performance during the past five years. When assessed for executive potential during the SESCDP selection process, the average candidate received superior (5s) or above average (4s) ratings in each of the nine evaluation criteria. Weaknesses, when noted, were generally in the interpersonal insight and problem analysis categories.

Pilch noted that younger candidates (less than 40) were less likely to be chosen for the SES program than their older counterparts. No adverse impact on either sex was observed; race information, on the other hand, was not available to make a similar statement in that area. On the whole, administrative managers were rated higher than scientific managers.

In summary, Pilch found that these managers can handle stress and articulate their thoughts clearly. While improvement in their leadership skills and

written communications was needed, the CAP participants were judged to be generally decisive and capable of adapting to a changing organizational milieu. Areas in which the managers most need to improve their skills include planning and organization, problem analysis, and interpersonal insight. For the latter two dimensions, in which over 65% of the evaluated managers exhibited deficiencies, developmental programs were contemplated to address those training needs.

True or False: The Assessment Center is an Organizational Panacea?

Jack Clancy

Clancy answered that question in his opening statement and then proceeded to explain what an assessment center is--and what it is not.

The assessment center is not a panacea for the ills of an organization. As Clancy noted, organizational structure and managerial needs must be considered before embarking on the assessment center or similar process. There is also no substitute for doing it the right way the first time--short cuts are eventually time-consuming and often very costly (i.e., litigation, etc.).

Clancy voiced his agreement with Brumback who noted earlier that job analysis is critical to the assessment center process because all dimensions to be evaluated should be operationally defined. The critical incident approach favored by Clancy lends itself better toward the development of assessment exercises than do other job analysis methods.

Since the assessment center is actually a supervisor/management performance test, the exercises to be used should reflect what a person in the target job actually does rather than their ease of administration. Clancy indicated that good exercises are often less than effective because of their scoring systems. An objective scoring system is a necessary requirement for every assessment center exercise. Such a system might be a five-point behaviorally-anchored scale with an operational definition and examples of observable behavior for each scale point. The obvious lesson Clancy noted here is that the more information you give the assessors, the more accurate the candidates' ratings will be.

Regarding assessor training, Clancy stressed quality rather than quantity (hours or days). He suggested that the purpose of the assessment center (e.g., development, selection, etc.) will often determine the training methods to be employed. In most cases, three days (20 hours) is the recommended minimum training time, and consulting costs, while generally necessary, can be reduced through the use of practice exercise videotapes and efficient scheduling by the user organization.

In summary, Clancy went on record as favoring assessment centers for most management development programs but urged caution in taking a similar universal approach on management selection issues. His presentation was followed by a brief question-and-answer period that dealt primarily with prescreening for assessment centers. Clancy felt an objectively scored in-basket test

offered the best prescreening solution; oppositely, Buck favored agency nominations to prescreen candidates, while Brumback suggested the use of a training and experience evaluation.

INVITED ADDRESS

Contemporary Personnel Psychology--An Overview

Dr. Thelma Hunt
Director, Center for Psychological Service
Washington, D.C.

In her presentation, Dr. Hunt reviewed the most important personnel concerns of today as well as discussing areas of needed or continuing research. She began, however, with an interesting historical perspective centered around her early efforts in test construction and validation. Harkening back to the mid-twenties, Dr. Hunt observed that the assessment problems of fifty years ago, test "standardization" and pre-test availability for example, have contemporary counterparts that are still being worked on in the eighties.

Addressing current issues, she highlighted validation problems, performance appraisals, employee involvement, oral board examinations, and passing points. In the validation area, Dr. Hunt does not view differential validity itself as a significant concept. Rather, she suggested that it is more important to equate test scores for racially, sexually, or ethnically different groups with equalities of job success. As such, if two different scores were judged equal in employee success prediction, the establishment of separate employment lists for various groups may indeed be justified.

Dr. Hunt acknowledged that tests are not perfect but noted that the fairness questions raised by this imperfection have social problem definitions which should be differentiated from the technical, psychometric explanations. Having observed that present-day definitions of validity are often court-supplied, Dr. Hunt encouraged more cooperative efforts between the legal and personnel professions.

Performance appraisal has been most often associated with criterion validity. However, with the recent escalation of civil service reforms, Dr. Hunt cited several other significant purposes for appraisal--namely, to insure retention of a competent work force and to serve as the basis for various personnel decisions. Both of these purposes, she felt, will demand more attention in the future.

Similarly, employee involvement has been recognized as an essential ingredient in personnel management. The method, according to Dr. Hunt, is not as important however as the recognition of its usefulness and necessity in establishing personnel decision procedures. Employee involvement is also seen as the first priority in oral examination development.

For oral examinations, Dr. Hunt indicated a preference for behaviorally-anchored categorical ratings but she cautioned that legislative edicts governing the determination of passing points often mitigate against such procedures. Hence, our role in personnel assessment should be directed more toward influencing change in outdated and inappropriate personnel laws.

On research for the future, Dr. Hunt stressed the need to make research statistics interpretative and understandable and to assure that research endeavors can be applied and utilized. In closing, she discussed the particular areas where research efforts should be intensified: motivation, leadership, work and human development, evaluation of personality characteristics and cognitive factors, longitudinal work history studies, and training.

PAPER SESSIONS

MEASUREMENT IN POLICE SETTINGS

Chair: B.J. Fuller, Consulting Services, Stansbury Park, Utah

Discussant: Tom Tyler, MEAS, Inc., Flossmoor, Illinois

Training Correlates of a Police Selection System

Ernest M. Johnson, Green Bay, Wisconsin
Michael A. McDaniel, Montgomery County, Maryland

A criterion validation study for a police officer selection system was presented. The selection components consisted of a written multiple-choice examination, physical agility test, and a structured oral board examination. Training academy measures were used as criteria (e.g., 13 written test items, markmanship ratings, and 24 performance evaluations across 12 dimensions).

Scores on the written selection examination significantly predicted scores taken on 13 written test items within the academy, $R = .74$. Scores on the physical agility selection examination significantly predicted 5 physical conditions ratings taken at the fourth week in the academy, $R = .57$, and at the twelfth week, $R = .45$. The structured oral board significantly predicted performance ratings on 8 dimensions taken during the final week of the academy, $R = .47$.

Differential impact of the selection system, using a breakdown by race and gender, was present in that minority applicants performed worse than white applicants on the written examination and males performed better than females on the physical agility measure.

The discussant emphasized the appropriate use of R rather than R^2 as the measure of a predictive relationship within a validation framework.

Physical Agility Tests That Result in No Adverse Impact

Matthew G. Forte, Port Authority of New York and New Jersey

The presentation described the development of a job-related physical agility screening device for police officers which did not include the usual calisthenic exercises. Based on three separate job analysis studies, the agility test consisted of a 150-yard run (including a four-foot wall, three-foot tunnel, zigzag course) eye-hand coordination test, 140-pound dummy drag, and 300-yard distance run.

The highlight of the presentation was the use of a computer program to identify cutting scores on each of the four components in the physical agility test to avoid adverse impact. The computer program produced all

possible combinations of cutting scores across the four components. Using the 80 percent rule of determining adverse impact, it identified those cutting scores which would not yield adverse impact across ethnic and sex groups.

Recommendations for the Use of Peer Rankings in Evaluation of Police Officer Performance

Kevin G. Love, Central Michigan University

The presentation discussed an alternative method of performance evaluation for patrol officers. Based on data indicating the suitability of the job of police officer for peer assessment (i.e., significant contact among squad members), an empirical study of peer rankings examined reliability, validity, and friendship bias.

The inter-rater reliability was significant, $r = .62$, as was the validity of peer rankings as compared to supervisor rankings, $R = .60$, and supervisor ratings, $R = .53$. Friendship as measured via ratings did not significantly affect the validity of the peer rankings.

The peer rankings were also found to be significantly related to the average number of on-job injuries, $r = .25$. This relationship may provide empirical support for the use of physical agility requirements for police officers in that superior police officers were exposed to significantly more potentially injurious situations.

The use of peer rankings in a comprehensive performance evaluation system for police officers was suggested. It was stated that peer rankings may provide utility as criteria in test validation studies.

FINE TUNING OF SELECTION INSTRUMENTS

Chair: David Friedland, Friedland Psychological Associates, Beverly Hills, California

Discussant: Jennifer French, San Bernardino County, California

Variably Weighted Distractors Used in a Parole/Probation Officer Written Exam

Grady Barnhill, Georgia State Examining Board

Job analysis indicated that counseling, resource utilization, recommending a course of action, and identifying needs and priorities were highly important areas of coverage and constituted some of the most important tasks performed by probation and parole officers. In order to assess these skills,

it was felt that a "case history" type of approach would be useful, presenting fairly detailed information about one probationer or parolee, and then asking a series of questions based on that information. This approach seemed to more accurately reflect actual job conditions than an approach which provided only a limited amount of information and presented only one question. During the item development sessions, it quickly became apparent that if questions were to be asked which were not trivial, obvious or irrelevant to some degree, the questions would need to address difficult issues and situations which probation and parole officers often face, and that oftentimes there would be more than one way to resolve a problem or situation. In the item development sessions, an effort was made to develop items which were "resolvable", as distinct from those casework situations which simply have no good solutions. However, no effort was made to shy away from those situations which are difficult to resolve and may not have "clear cut" solutions. Field experts were asked to bring case histories of actual parolees or probationers or case histories based on actual events. The incumbents were provided with a brief summary of principles to observe while constructing items, processes used in the generation of the areas of coverage were discussed, and different field experts were asked to develop items in specific areas of coverage. In order to most accurately reproduce job task conditions, some items were based upon detailed case histories which included such information as might be gleaned from interviews with family members, a perusal of police records, and other investigative procedures. In order to assess an applicant's ability to utilize resources, a resource booklet was prepared which described various facilities, institutions or persons similar to those which a probation or parole officer may use in the course of his or her duties. Some questions were developed which require an applicant to evaluate and consider the information presented in the case history and make decisions as to which of the described resources is appropriate to use. In the final assembly of the test, questions were arranged in such a manner as to provide cumulative information. For example, an applicant may need to remember information from a conversation described in one question in order to answer a following question. Oftentimes it proved difficult to create an item which assessed skills in only one area. Typically, an item assessed skills in several different areas of coverage at once. This was apparently true because of the interrelated nature of the areas of coverage. Obviously, it is difficult to perform counseling without utilizing decision making skills, and it is impossible to make the best use of resources without using both counseling and decision making skills.

Field experts generated and reviewed items. At several joint item development sessions, the items generated by field experts and by test development staff were reviewed for technical accuracy, realism and appropriateness, and then a draft of revised items was typed for a more systematic item review. A group of 22 probation and parole officers, some of whom were at the supervisory level, was assembled to evaluate the items which had been generated. The group was overrepresentative with regard to minorities and females, as had been the case in job analysis sessions. First, the field experts were asked to look at each item and record the answer they felt was most appropriate on an answer sheet. Secondly, the incumbents were

asked to evaluate each item, indicating which area of coverage, if any, the item related to, the difficulty of the item, the relationship of the item to the job, and the degree to which the item would distinguish between adequate, inadequate, and superior performance on the job. In addition to this information, the field experts were asked to evaluate the relative worth of each choice of each item. Each choice was assigned a value from "1" (very poor) to "5" (very good) by each field expert.

Due to the nature of the items, it was determined that the most appropriate scoring procedure would be the use of variably weighted distractors. The mean value assigned to each distractor by the final group of 25 field experts was the starting point for determining what value to award for each question choice. The standard error of measurement was added to and subtracted from each mean value of each question to establish a "range" for each mean. If two or more means in one question had ranges that overlapped, then those means were considered to be equivalent and were averaged together. If three or more of the means in one question were averaged together using this criterion, then it would be difficult to obtain the kind of "multiple level" discrimination referred to earlier, and the item was discarded. If the mean value for each question added to its own standard error of measurement did not equal at least 2.5, then that choice was not given any credit. Since 3 was the value for a "good" response and 2 was the value for a "poor" response, it was felt that a 2.5 cutoff point would prevent crediting any response which was more "poor" than "average." Additionally, this was done to help provide maximal discrimination between the poor and the average candidate.

The number of times that field experts selected a choice also played a role in determining the distractor weight for each question. In order to eliminate questions which might be too confusing even for the field experts, it was decided that if at least 50% of the field experts did not agree on the best answer, the question would be discarded. In order to increase the ability of the test to discriminate between a person who selects a best answer and one who selects a second, third, or fourth best answer, it was decided to increase the mean value of the distractor by a specified percentage. The more field experts agreed on the best answer, the larger was the percentage increase in the mean value. Since the more clear-cut questions would have a higher distractor value, a person missing that choice would, in effect, lose a larger number of points than the person who missed a more subtle question. Typically, the choice selected most often for a particular question by the field experts was the same choice which was assigned the highest mean value by the field experts. If the average "distinction" value for an item was less than 1.75, the item was discarded. A value of 1 on this scale represented an item which was not likely to make a significant distinction between levels of competency, while a value of 2 indicated that the item was likely to distinguish between adequate and inadequate levels of performance. If the average "relatedness" value assigned to an item was less than 2, then the item was eliminated.

In reviewing the operation of the test, it appears that the procedure of weighting more heavily those items which were more clear-cut (upon which

there was more field expert agreement) may be causing raw scores to be closer together than is desirable. While this procedure was designed to distinguish between the adequate and inadequate applicant, it appears that these questions are indeed easier for everyone to answer correctly, and thus, a number of raw score points are being added to most testee's scores. These higher raw scores have a tendency to mask the discriminating power of the test in regard to the items which distinguish between the adequate and the superior applicant. Since the differences in point values for the more difficult questions are smaller, these small numerical differences can be easily obscured by the addition of the large number of raw score points that applicants receive by correctly answering easier questions. One possible method of correcting this tendency would be to eliminate any consistent weighting of choices on those questions where only one distractor is to receive credit according to the criteria described above. The easiest questions have only one distractor which receives credit, and eliminating the extra weighting of these questions would help correct the problem described above.

Detection of Test Item Bias and Its Effects on Group Test Performance

Darryl Lang, industrial/organizational psychologist, Denver

In the present study, an item bias analysis was carried out based on latent trait theory. In general, this test bias method is based on each group's probability of responding correctly to an item, which is computed with ability level taken into account.

A very desirable property of latent trait theory is that the shape of the item characteristic curve does not vary across subgroups of examinees from the examinee population. In other words, the estimated item parameters and person abilities are not dependent on the ability distribution of the examinee sample. With classical test theory, however, item and person parameters are not stable across samples with different ability distributions. This property, called parameter invariance, is an important advantage of latent trait models.

The one-parameter latent trait model, based on the work of Georg Rasch, a Danish mathematician (Rasch, 1966), was used in the present study to detect biased test items. In general, the Rasch model describes the probability of a person's success on an item as a function of the person's position on the trait or ability and the difficulty of the item--and nothing else. The model assumes items have equal discriminating power and vary in difficulty only. The Rasch model was chosen as the method of item bias detection in the present study for the following reasons:

1. The Rasch model item parameter, difficulty, has been shown empirically to be efficiently and consistently estimated from observed item responses (Andersen, 1973; Wright and Douglas, 1977). Other latent trait models result in additional item parameters that have not been as efficiently and consistently estimated.

2. The one-parameter model gives stable estimates with as few as 100 examinees (Wright, 1977). A large sample size (approximately 500 or more) is required to derive stable parameter estimates with the three-parameter model--a condition not met in the present study.

3. The Rasch model makes strong assumptions about the nature of test data; and if these assumptions are not met, Rasch measurement characteristics (e.g., parameter invariance) cannot be implied. Thus, it is important to detect persons and items that do not fit the model. Statistical procedures have been developed by Wright and Mead (1977) to determine if test data fit the Rasch model. Similar fit statistics have not been extensively developed for other latent trait models.

A 99-item multiple-choice fire engineer promotional exam was designed as a screening device for the promotion of fire fighters to the position of fire engineer in a large midwestern fire department. The test, based on a fire engineer job analysis, contains four content areas: General, Fire Fighting, Preventive Maintenance, and Safety. Each test item has four alternatives and was scored dichotomously.

A principal factor analysis was carried out to determine the dimensionality of the promotional exam. The results of the factor analysis did not show the test to be multidimensional. Thus, the assumption of unidimensionality required by the Rasch model was met.

The fire engineer promotional exam was administered to a group of fire fighters (N = 1,038). Eight hundred and ninety-one whites and 147 blacks, all male, were in the group. A classical item analysis was carried out on the test items. Items with point-biserial correlations (item-total) less than .20 were eliminated. Then, items with difficulties less than .3 and greater than .7 were eliminated. (This classical difficulty parameter is the proportion in the sample answering the item correctly and should not be confused with Rasch estimated item difficulties.) Thirty-five items were eliminated, and the remaining 64 items met the criteria for "good" items. These items were designated as the "classical item set."

A Rasch fit analysis was carried out on the promotional exam. A computer program first designed by Wright and Mead called BICAL (version three) was utilized to carry out both the person and item fit analyses. A person fit t-statistic is computed by the program internally. None of the person fit t-values exceeded the criterion for person deletion, and thus, no examinees were eliminated. Forty-five items were identified as not fitting the Rasch model. The remaining 54 fit items were computed on three groups: the black group and two white groups. The white group was divided into two subgroups to check on the reliability of the bias analysis. A difficulty shift t-statistic utilized by Draba was used to detect biased items. Significant difficulty shifts would indicate that there is a response-by-group interaction and possible racial or cultural bias. The 98-percent confidence

level was the criterion for significance. The t-values for each item were computed. First, shifts in item difficulties were computed between the two white groups to determine the method's reliability. Since the racial-cultural composition of these two groups is assumed to be the same, item difficulties should not be significantly different. As predicted, none of the items had significant t-values.

Next, the estimated item difficulties computed on the black group were compared to the estimated difficulties computed on the two white groups. Nine items had significant t-values for one or both of the group comparisons and were defined as "biased" items. Three items were significantly more difficult for the white group than the black group and the six other items were significantly more difficult for the black group. The remaining 45 items were called the "unbiased item set."

Besides a significant difficulty shift, group response differences could be reflected in an item not measuring the variable of interest for one of the groups. That is, an item or set of items does not fit the Rasch model for both groups. In the present study, all the fit items computed on the white group fit the model, but these same items computed on the black group did not. Seven items were identified as not fitting the model for the black group. The remaining 47 items were designated as the "black fit item set." Adverse impact was investigated by computing the proportion of white and black fire fighters who scored 2, 1.5, and 1 standard deviations above the total group mean on these item sets: classical, fit, unbiased, and black fit. The proportions for each group were compared and adverse impact was determined by using the "80 percent rule" specified in the Uniform Guidelines on Employee Selection Procedures. Adverse impact was also investigated using the Rasch estimated abilities for the fit, unbiased, and black fit item sets. A z-statistic was used to test the statistical significance of the mean test score and ability differences between the black and white groups. For all item sets, the white fire fighters' test scores and abilities averaged significantly higher than the black fire fighters.

Following the "80 percent rule," adverse impact existed for all item sets, for both total test scores and Rasch abilities, and at all three cut-points--2, 1.5, and 1 standard deviations above the combined group means. Compared to the classical item set, the Rasch item sets slightly lessened the effects of the adverse impact for the black group based on total test scores. Adverse impact was lessened slightly with the unbiased item set, placing five more black examinees at the 1 and 1.5 cut-points. This increase did not coincide with an increase in the white examinees scoring at both cut-points. In fact, the proportion of black examinees divided by the proportion of white examinees was slightly greater (closer to 80%) at both cut-points for the unbiased item set than the classical and fit item sets. The black fit item set had the greatest overall effect on the reduction of adverse impact.

The elimination of the biased items did not get rid of black and white mean differences with test scores and Rasch abilities. One reason is that the nine biased items were relatively easy items for both groups and had little impact on total test scores and Rasch abilities.

Overall, the results of the present study did not warrant any conclusions about the usefulness of the Rasch model either for detecting biased items or as an item analysis method for eliminating group score differences.

It should be pointed out, however, that in the present study and other studies the utility of the Rasch model is evaluated in comparison to classical test theory (e.g., reliability and validity); this may be inappropriate. The Rasch model and other latent trait models are presenting a new method of test development based on theoretical assumptions and conditions quite different from classifcal test theory. Theoretically, latent trait theory is an improvement over classical test theory and perhaps should be evaluated within its own theoretical framework.

Just how Good Are Foreign Credentials--A Model for Evaluating Quality of Training

Peggy Goulding, Goulding, Martin and Associates (presented by Charles Martin)

This project arose out of a court case involving the evaluation of foreign credentials. The area, especially of interest to consulates, embassies, and programs of the U.S. State Department, as well as to any employer of foreign-trained staff, involves issues normally associated with Training and Experience Evaluation plus some additional logistical complexities.

The first step involved job analysis to determine the job responsibilities and the necessary Knowledges, Skills, and Abilities (KSAs). The next step was to determine how an applicant might indicate possession of the prerequisite KSAs; so far, a standard T&E methodological process.

In the case of the foreign credentials, however, the third step was a determination of how foreign institutions train students. The mechanics involved discussions with faculty in the U.S. and in the particular countries of interest and discussions with graduates of foreign universities who were working in fields of interest in the U.S. to design a system of assessing the actual knowledge possessed.

This assessment system utilized a Subject Matter Expert (SME) panel of academicians and practicing engineers. Each applicant was rated by this panel on each dimension that was determined to be important in the job assessment. This rating was done in a manner to insure the "blind" application of the same standards to all candidates.

A conclusion derived from the study is that universities may need to augment efforts to document the salient learning outcomes of the various degree programs.

SYMPOSIUM

Unassembled Examining--

Current Approaches in the Federal Government

Moderator: Marianne Bays, Eastern Region, U.S. Office of Personnel Management

Presenters: Lou Dunn, Examination Methods Development Unit, U.S. Office of Personnel Management
Olivia White, Examination Methods Development Unit, U.S. Office of Personnel Management
Steve Norton, Department of Defense

Discussant: Steve Bemis, Information Science, Inc.

This session opened with a brief description of the traditional training and experience (T&E) rating process and the problems and disadvantages associated with it.

Mr. Dunn then explained why the Behavioral Consistency Model developed by Dr. Frank Schmidt of OPM's Personnel Research and Development Center is a significant improvement over the traditional T&E process. Some of the reasons presented include:

a) the BCM assesses quality based on achievement rather than merely exposure;
b) the predictive value is higher;
c) the BCM is less likely to have adverse impact since achievements are not situationally limited;
d) the rating is based on achievement which may have occurred in a wide variety of settings.

This model is not, however, without drawbacks. Typically, it takes 4-6 months to develop a rating schedule and implementation/administration costs may be slightly higher. Also, since the Behavioral Consistency Model was developed specifically to screen for mid-level positions, some problems were encountered in adapting the process to entry-level positions.

The rest of this presentation summarized the process (and some problems encountered) used to develop a BCM-based rating schedule for entry-level positions.

Next, Ms. White addressed the question of unpaid work in relation to unassembled examinations. After a brief discussion of some of the problems and inconsistencies typically associated with scoring unpaid work using either a traditional T&E or a BCM-based rating plan, Ms. White summarized a research project designed to:

a) identify ways to obtain adequate descriptions of unpaid work;
b) identify a way to obtain consistent and appropriate ratings for unpaid work;
c) develop a model crediting plan with benchmarks at each level for unpaid work, paid work, and academic backgrounds.

The project resulted in a number of usable benchmarks for unpaid work, a training package for raters, and a response collection methodology.

The final presentation in this session summarized a Department of Defense project for developing an unassembled exam process for selecting Quality Assurance Trainees. Due to the large number of applications to be handled, this was to be an automated process. Included in this presentation were examples of operationally defined KSAs generated during the job analysis process and sample items used to screen them.

SYMPOSIUM

The Structured Interview--Dead or Alive?

Moderator: Louis M. Laguardia, Personnel Psychologist, U.S. Office of Personnel Management, Eastern Region

Presenters: Philip Ferrara, New York State Office of Court Administration
Steve Nettles, Personnel Psychologist, Educational Testing Service (for Dr. Richard Thornton)

Discussant: Jerry Durovic, New York State Department of Civil Service

The moderator opened the session by expressing a general concern about the poor reputation which oral interviews have had in the past, in particular for their lack of validity and reliability. The moderator commented that this poor reputation sometimes is well deserved, but frequently is very well misplaced. He asked those who use structured interviews as a selection procedure not to swiftly blame poor and/or adverse results on the interview itself without first assessing if all the steps have been taken to insure the technical soundness and standardization of the procedure.

The moderator further indicated that there are primarily two points of view about structured interviews: On the one hand, there are those who consider the structured interview too unreliable and have little hope that anything constructive can be done about it. On the other hand, there are those who recognize the weaknesses of the structured interview, but take a more positive outlook by constructively directing their efforts toward improving the procedure.

The moderator ended his remarks by emphasizing that the objective to pursue was to maximize the validity and unreliability of the structured interview through means already known, without having to seek the total elimination of subjectivity in the procedure, a task he considered logically impossible.

Steve Nettles presented a paper in place of Dr. Richard Thornton entitled "Developing and Designing Reliable Standardized Interviews." Mr. Nettles reported on a project funded by the U.S. Department of Labor in which selection programs were developed for the positions of Interviewers and Local Office Managers. The selection measures for both positions included written, objectively scored tests and standardized interviews designed to assess tasks and/or knowledges, skills and abilities.

In order to implement the program in all fifty states, the Department of Labor funded a series of training programs for representatives for each of the states in the use of the instruments. Mr. Nettles focused his presentation on those aspects and results of that training program that concerned the standarized interview procedures developed for both positions.

Among the issues discussed, emphasis was made that an important aspect of conducting a standardized interview is to prepare in advance a series of questions that will both limit the areas to be evaluated to their specific aspects of the job most appropriate for evaluation by means of an interview and will provide an opportunity for candidates to present information relative to their qualifications in the areas. Mr. Nettles explained one or more questions were prepared for each of the five responsibilities to be assessed, and the questions were reviewed by job experts to insure their reasonableness and relationship to the responsibilities.

As far as standards for rating candidate responses, specific examples (behavioral anchors) for the three-interval rating scale were prepared by job experts. The three intervals comprised a scale ranging from an unacceptable response through an acceptable one to a superior response. The three intervals were also assigned numerical values so that a combined score could be obtained by averaging the ratings given the responses for each of the questions. The total score on the interview was the average of all three raters. While raters could vary their ratings of candidates between acceptable and superior, in the case of an unacceptable rating, however, the procedures of the interview required that it be the unanimous rating of all three panel members.

According to Mr. Nettles, the procedures were pilot tested using four interview boards in two states. A total of fourteen candidates, all eligible to compete for the position of Local Office Manager, were interviewed. The procedures proved effective, and the results were highly reliable in this small sample.

Mr. Nettles also discussed at length the steps taken to standarize the training for panel members in all fifty states. Audio-Visual Simulation techniques were used for this purpose.

The results reported were as follow: In all, 54 panels of interviewers observed and scored the three simulated interviews. Intraclass reliability coefficients were computed for each panel. The lowest coefficient obtained was .69. Ninety percent (90%) of the reliability coefficients exceeded .85, and 80% exceeded .90. The ratings of the three interviews were remarkably similar over all groups with one interview marginally acceptable, one clearly acceptable, and one consistently rated superior.

In summary, Mr. Nettles stated these results seem to confirm that, when steps are taken to standardize the content of interviews and the method of evaluating or rating the results is carefully scaled, the scores or results of interviews can be quite reliable. It was noted that the interviews in this study were simulated in order to reduce the variance that would be encountered should different interviewees be employed or even the same interviewees subjected to repeated assessments.

Mr. Nettles reported that in the opinion of the author (Dr. Richard Thornton) the results reported in this presentation more accurately reflect the reliability of the procedures described as they would be used in a selection situation.

In the second presentation, Dr. Philip Ferrara described the competitive selection procedures that are in use by a large state court system for the hiring and placement of Spanish-speaking Court Interpreters. The two-stage selection process consists of a written test followed by a structured oral interview.

According to Dr. Ferrara, the written test portion presents the candidates with a series of tape recorded statements dealing with subject matter relating to the Civil, Criminal, or Family Court setting. Candidates are required to select the correct translation in Spanish of materials recorded on tapes in English or vice versa. Top scoring candidates are then administered the structured oral interview which places the candidates in a courtroom setting with a simulated mock trial.

Dr. Ferrara presented detailed information about the procedures employed in: the selection process for bilingual oral examiners and live actors; the development and standardization of oral exam scripts that have been derived from actual court transcripts and task-based job analysis information; the development of behaviorally-anchored rating scales; the examiner training session; and the candidate appeal process.

Dr. Ferrara highlighted the advantages associated with this technique over the "conventional" oral interview from the perspective of job relatedness, selection procedure standardization, and legal defensibility. In the final section of this report, Dr. Ferrara addressed how the public as well as the private employer can intelligently and fairly approach personnel selection and placement when faced with non-English speaking or bilingual job applicants.

Some of the results Dr. Ferrara shared in his presentation were as follows: Out of a total of 140 candidates who went through the structured oral examination, 118 candidates were successful. Candidates were fairly well distributed across the range of passing scores. Not correcting for restriction of range, the pre-screening written examination correlated .40 with the structured oral. Dr. Ferrara explained that, in order to assess the reliability of the ratings, intraclass correlations were computed using an analysis of variance paradigm as proposed as Ebel (1954). The results obtained according to Dr. Ferrara reflect the fact that raters can make accurate decisions when there are specific criteria and an opportunity to observe actual behavior that is expected in the target position.

PAPER SESSIONS

ISSUES IN JOB ANALYSIS

Chair: Doris Maye, Georgia

Discussant: Charley Sproule, Pennsylvania State Civil Service Commission

Classification and Testing: An Integration

Reginald A.H. Goodfellow, California State University

The functions of classification and pay and testing and selection are all too often treated as separate functions by many organizations. This separation not only results in increased costs but ignores the fact that both functions have as their point of origin the same source--the job.

A study was performed for a large Southern California public agency and an integrated "Job Information Base" (JIB) was developed which was especially designed to provide a systematic technique for linking the tasks performed and the personal characteristics (SKAOs) required for performance. The ultimate goal was a data composite directly useful as . ut into classification, recruitment, selection, and evaluation.

The JIB process begins with a description of the tasks or duties performed in a position. This description may vary in a number of ways, but one most important dimension is that of the detail involved in the description.

For example, a task statement for a teacher may read:

> Writes important lesson points on blackboard.

Another view of this task may read:

> Evaluates lesson content. Decides which points will highlight lesson. Uses chalk to write the outline legibly on board. Verbalizes to class the importance of the outline. Ensures that students understand function of lesson outline.

The second description is much more detailed than the first and provides a rich source of information about some of the personal characteristics which are required of the teacher.

The description couched in somewhat molar, or general, terms is the "task," while the more molecular description of the smaller "tasks" involved in the above example are the "behaviors." A behavior here is defined as something which can be observed or reported as occurring during the performance of a task.

A detailed analysis of the "behaviors" associated with the tasks performed on the job provides the basis for the JIB method of determining and linking the SKAOs with the duties of the position.

For the generation of behaviors, the JIB process assembles a group of incumbents and/or supervisors in a workshop format and asks them to provide a list of tasks or duties performed in the position. This list is worked on until the workshop participants are comfortable that all of the tasks have been described. Then, using each task in sequence, the workshop participants are requested to describe the "behaviors" associated with each task. These behaviors are listed singly on 3x5-inch cards. Each task is then examined until the list is exhausted. The net result is a large number of index cards, each containing a discrete behavioral description which is tied to some job activity. A procedure called "content analysis" is then used by job analysts to make sense of the behaviors.

The major activity in a content analysis is one of sorting and classifying responses into distinct, identifiable response categories, or performance dimensions. It is a time-consuming and difficult endeavor. The difficulty of this technique lies in the inherent complexity of any judgmental process which calls for a sorting operation based upon common elements. There are apt to be several response categories which are, at once, conceptually related but discriminably different. The problem is one of deciding whether some particular distinction is worth preserving. If several job analysts are to be involved in the process, there is some merit in first doing a content analysis of their own notes independently. When this is accomplished, they might come together and in a joint effort go through a final sorting process of the combined data. The end products of a content analysis are stacks of cards, each containing several statements relevant to a single concept of behavior, each stack representing a performance dimension or SKAO.

As an example, some of the behaviors defining an "Evaluation of Academic Progress" dimension might be:

Develops minimum standards for academic achievement

Designs oral and/or written questions to test students' understanding of subject matter

Sets up consistent criteria for evaluation of academic progress

Devises inferential questions to test for understanding of material

Explains academic evaluation procedures to students

When the job analysis has proceeded to the point of identifying the tasks and SKAOs, a major portion of the work is complete. The next step involves verifying the data collected and establishing, insofar as possible, the relative importance of the tasks and SKAOs. This stage is performed by creating a questionnaire or questionnaires which request all incumbents (using a population eliminates "sampling" problems") to rate:

(1) The criticality and frequency of tasks

(2) The criticality and frequency of behaviors

(3) The criticality and frequency of overall performance dimensions

Each frequency rating is accomplished by using a seven-point scale with descriptive anchors at each level--1 reading "At least once but no more than a few times a year" and 7 reading "Continuously." The criticality ratings also utilize a seven-point scale with descriptive anchors at values 1, 3, 5 and 7. For this rating 1 reads "This task is of minimal importance to the job and even if performed poorly would not have much overall effect on the successful performance of the job" and 7 reads "This task is of critical importance to job success and poor performance will always prevent the job from being accomplished successfully."

Results are analyzed and means, standard deviations and frequency counts are obtained for each variable. The reliability of each performance dimension is assessed by coefficient alpha derived from the subsets of behaviors which define the content of a particular performance dimension.

The final result of the job analysis phase is a set of very detailed data which can be used to develop special application forms, interviews, oral examinations, performance appraisals, job descriptions, job specifications, etc. Also the availability of detailed questionnaires makes it relatively simple to assess the usefulness of the data for other organizations and to assess job changes which may occur over time.

The Measurement of Job Similarity for Test Transferability

Harold Bartlett and Arthur Rosenblum, Personnel Assessment Corp., Denver
(presented by Harold Bartlett)

There are numerous public jurisdictions which do not have the necessary resources to validate their own selection systems and examinations. The Uniform Guidelines on Employee Selection Procedures endorse the concept of transferability of validated tests provided, however, that it can be shown that the job for which the test was developed is substantially the same as the job to which the test is to be applied. Other than making reference to the comparison of job analyses, the Guidelines do not provide a methodology for establishing job similarity. The concept of test transferability implies two courses of action: (1) Transferred Validity in which one jurisdiction borrows the test validated elsewhere, and (2) Cooperative Validation in which two or more jurisdictions pool their resources in a validation study. In either case it is necessary to establish a methodology to determine the extent of the similarity of jobs in the different jurisdictions. Following is the methodology for test transferability developed for the Physical Performance Test used by the Denver and Lakewood, Colorado, Fire Departments.

The entire procedure depends upon the existence of competent job analyses which contain a list of the task elements and knowledges, skills and abilities (KSAs) needed to perform the tasks. Equally necessary is a measure of criticality for each of the tasks and KSAs.

With their measure of criticality as a data base, the task lists from each jurisdiction and the KSA lists from each jurisdiction are subjected to a measure of similarity by means of a Profile Similarity Coefficient developed by Cattell, r_p. This coefficient has range characteristics similar to that of the Pearson's correlation coefficient; that is, -1 to +1, with zero indicating no relationship. Appropriateness of transferability of test cutoff points can be determined by using the Profile Similarity Coefficient to compare difficulty measures, if available, of the tasks or KSAs from the two job analyses. If difficulty measures are not available, rational justification of the cutoff may be utilized.

The benefits of r_p as a job comparison method are: (1) it is a positive measure of the similarity of jobs rather than only the difference; (2) it is subject to a test of statistical significance; (3) it can be used to compare as few as two jobs or as many, pairwise, as desired; (4) it does not suffer from the weaknesses of logic inherent in a null hypothesis testing approach; (5) it is easy to calculate in the uncorrelated elements form; and (6) where uncorrelated elements cannot be assumed or created, corrections for obliquity can be made.

A summary of the procedure itself follows:

1. Obtain and rate lists of job elements and KSAs for importance and difficulty in the jobs to be compared.
2. Multiply importance rating by difficulty rating to obtain overall rating.
3. Initially, calculate the pooled mean and standard deviation of rating scores across raters and elements and KSAs.
4. Convert ratings to standard scores and calculate the difference in scores between mean ratings on each job.
5. If elements can be assumed uncorrelated, calculate r_p as in equation (1).
6. Compare with Horn's r_p significance table.
7. If elements cannot be assumed uncorrelated, find the correlation matrix among elements by correlating across raters. (This is not legitimate unless the total number of raters exceeds the number of elements.)
8. (a) Carry out a components analysis and simple structure rotation on this correlation matrix.
9. (a) Calculate factor scores for the raters using equations (2) or (3), whichever is more computationally convenient.
10. (a) Calculate the mean factor scores on each job across raters for that job and calculate the difference scores between mean factor scores on each job.
11. (a) Calculate r_p as in equation (1) and compare with Horn's table for significance.

or

8. (b) Do a components analysis on the correlation matrix in Step 7 to obtain Λ the diagonal matrix of its roots (eigen values).
9. (b) Calculate r_p as in equation (1) and compare with Horn's significance tables.

The question of how substantially the jobs must be the same is, of course, unanswerable. The personnel specialist can, using this method, discover the extent of relationship and whether or not that relationship is greater than what would be expected of a chance relationship. This state of affairs is analogous to the reporting of the correlation coefficient between two random variables such as test scores and job performance. Judges or experienced personnel specialists may make decisions about what level of association is acceptable that they could not make before such a measure was in use. For jobs meeting this decisional criterion, tests may then be transferred. For groups of jobs that mutually meet the decisional criterion, cooperative validation is a reasonable procedure.

Final documentation of the validity of test transferability should at least include a side-by-side presentation of the two pairs of lists, rank ordered by their criticality measures for comparison, similarity coefficients for each pair of lists along with their respective levels of statistical significance and, lastly, a descriptive narrative covering job similarities, differences and difficulty levels.

Equations:

(1) $$r_p = \frac{2k - \sum d_j^2}{2k + \sum d_j^2}$$

where k is the median χ^2 value for degrees of freedom equal to the number of profile elements and d_j is the difference of standard scores between the two jobs on profile element j. The summation is across all profile elements.

(2) $$F = (V'V)^{-1}V'Z$$

The matrix of factor scores F is equal to the inverse of the matrix formed by premultiplying the factor (pattern) matrix V its own transpose V' multiplied by the transpose and finally post multiplied by Z, the observed standardized scores.

(3) $$F = V'P\Lambda^{-2}P'Z$$

Where F is the matrix of factor scores desired, V is the rotated factor matrix as before (e.g., a varimax solution). P is the original principal axis solution. Λ^{-2} is the diagonal matrix containing $1/\lambda_i^2$ in the diagonals where λ_i is the eigenvalue corresponding to factor i and Z is the matrix of observed scores in standard (Z - score) form.

Job Evaluation and Wage Comparability: The EEO Issue of the 1980s

Lance Seberhagen, Seberhagen and Associates, Vienna, Virginia

Women have traditionally earned about 60% of what men have earned, and a similar pattern exists for minorities in comparison to the majority group. The exact causes for these differences in pay are difficult to isolate. Salaries are determined generally on the basis of the intrinsic worth of the job and the relative bargaining power of the parties concerned. Intrinsic worth is usually measured through some sort of job evaluation procedure, while bargaining power is based on a combination of factors such as supply and demand, ability to pay, work performance, unionization, negotiation skill, political power, and the militancy of each party. Employment discrimination could affect both intrinsic worth and bargaining power.

The Equal Pay Act of 1963 has prohibited employers from paying different salaries to men and women who perform substantially the same job, unless such pay differences are due to seniority, work performance, or some other factor other than sex. Legal authorities agree that Title VII provides at least the same protections as the Equal Pay Act but covers not only sex but also race, color, religion, and national origin as well. The real legal controversy is whether Title VII requires "equal pay for work of equal value." In other words, should jobs traditionally held by women (e.g., nurse) be paid the same as jobs of "comparable worth" which are traditionally held by men (e.g., plumber)?

Until the early 1980s, the EEOC had done little in the area of wage discrimination. In April 1980 the EEOC held public hearings on "Job Segregation and Wage Discrimination" and announced that it was going to step up its efforts in this area. Chair Norton said that these hearings were the most important that EEOC had held in a decade and that she expected wage discrimination to be one of EEOC's top priorities for the 1980s. Shortly after EEOC's hearings on wage discrimination, it was revealed that the National Academy of Sciences had drafted a preliminary set of guidelines on job evaluation for EEOC. These guidelines do not require all employers to use job evaluation but are intended to provide direction to those who do. The major provisions of these guidelines are:

> Employers may tailor job evaluation methods to meet the individual needs of the organization, but job evaluation methods which create an adverse impact are illegal unless they can be shown to be a business necessity;
>
> Employers should adopt either one job evaluation method or a set of interrelated methods so that all jobs are essentially rated against the same standards of worth;
>
> The employer's concept of job worth should be an explicit and open policy of the organization;
>
> Job evaluation factors should measure all important aspects of the employer's concept of job worth, without contamination from other, irrelevant considerations;

The technical details of job evaluation procedures should be well documented and readily accessible to all employees;

Formal steps should be taken to eliminate sex or race stereotyping in the design and administration of the job evaluation system;

The job evaluation system should contain an internal appeals mechanism;

All minimum qualifications for jobs should be valid;

All job titles should be standard for positions having essentially the same duties.

One way to investigate possible sources of discrimination is the development of multiple regression models to isolate the effects of each variable. Such a study, based upon a random sample of 301 (192 men and 109 women) full-time, permanent state employees, revealed that the median women's salary was about 70% that of men's and the mean women's salary was about 67% that of men's. The size of these salary differences is somewhat less than for the general workforce but fairly typical for the public sector.

Based on earlier research by the Institute for Social Research at the University of Michigan, a preliminary set of merit predictors of salary was established which consisted of occupational prestige (a type of job evaluation measure), education, hours worked per week, number of employees supervised, total state tenure, and position tenure. Separate regression equations for men and women using these predictors produced a multiple correlation, adjusted for shrinkage, of .81 for men and .74 for women, both of which are statistically and practically significant. Two attempts to improve on this level of prediction were unsuccessful. The first attempt added scores from the Wonderlic Personnel Test and the Achievement and Self-Confidence scales of the Gough Adjective Checklist. The second attempt added ratings of the relative importance of five job factors: comfort, challenge, pay, coworker relations, and adequacy of resources, as measured by a 23-item questionnaire used in the University of Michigan research. Thus, it was concluded that the point of diminishing returns in the prediction of salary on the basis of merit-type variables had been reached.

The next step was to apply the original male regression equation to women and the original female regression equation to men. The difference between a person's actual salary versus his/her predicted salary if he/she were the opposite sex gives one possible indicator of wage discrimination. When the male model was applied to women, women's mean predicted salary increased about 34% (from $575.48/month to $771.99/month at 1973 salary levels) to the point where women now earned about 90% of what men earned ($857.02/month). In other words, only about 30% of the original salary difference could be explained by differences in merit, and the remaining 70% of the difference is probably due to sex discrimination. Similar results were found when the women's salary model was applied to men.

A corollary analysis of the data combined the total sample of men and women in one regression equation using the original merit predictors plus sex

as a dummy variable. After correction for shrinkage, sex alone correlated .48 with salary, and merit alone had a multiple correlation of .77. The addition of sex to the merit predictors increased the multiple correlation significantly to .84. Thus, sex alone accounted for 23% of the variance in salaries, and merit alone accounted for 59% of the variance in salaries. When merit was controlled, sex still accounted for a significant 11% of the variance in salaries, leaving 30% of the variance due to other variables not specified. Based on the regression weights for the full sample, sex accounted for 71% of the dollar difference in mean salaries between men and women after merit had been controlled.

The practical implication of these findings is that while some of the gross differences in pay between men and women may be due to legitimate factors such as merit, about 70% of the difference in pay seems to be strongly due to sex.

IPMAAC should be concerned about wage discrimination for a number of reasons: (1) "assessment" methodology applies to job evaluation as well as to employment testing; (2) EEOC may develop new guidelines on job evaluation and position classification which could affect job analysis and testing procedures; (3) no other section of IPMA has the organization or expertise to deal with these issues now; and (4) IPMAAC should keep a broad definition of its charter to maintain its vitality.

TEST VALIDITY AND UTILITY

Chair: Richard Hodapp, Wyoming

Discussant: Les Canges, Colorado

Criterion-Related Validation of Tenure Predictors

Michele Fraser, Richard D. Olsen, Lowell Hellervik, Marvin Dunnette, Personnel Decisions, Inc., Minneapolis

In positions involving training, tenure is as important a criterion of job success as job performance. This strategy was developed for a management trainee position.

The researchers used a rational approach based on the job analysis and their ideas about what affected job tenure. They came up with five subscales:

1. Level of work motivation
2. Level or work energy and work pace
3. Compatability of personal values with job characteristics

4. Past and expected enjoyment of specific work activities
5. Tolerance for negative aspects of the work

They constructed biographical data scales to measure the five subscales. Turnover literature supports the idea that the single best predictor of turnover is intention to stay or leave. It was decided to ask people how long they would stay and correlate subscales with their report in order to corroborate the usefulness of the subscales. Employees were also asked to give the odds of their staying given lengths of time. The results were taken to be the longest time they reported at least 50% odds of staying.

Correlations of about .34 were found between intentions to stay and the five subscales. However, when data of actual tenure were analyzed, it was found that the correlations between intention and actual tenure was .16 and between the subscales and actual tenure was .11. The researchers speculated that the criterion used for this result (currently still employed or not) was insufficient for the evaluation. They are still confident that intention is a workable predictor.

Validating Performance Appraisal Forms--An Assessment of Quality Measures

John G. Veres III, Hubert S. Feild, Wiley R. Boyles, Auburn University at Montgomery

In a study of research articles in the Journal of Applied Psychology, it was found that fully 72% of validation studies reported use of performance appraisal forms.

The researchers decided to look at typical operational performance ratings compared to more objective measures of job performance for clerical positions. The two more objective measures were actual tests of filing and proof reading skill and specially designed research ratings involving more specific behaviorally-anchored scales and rater training for their use.

Findings were that research ratings were less lenient than the operational performance appraisal ratings and there was somewhat less central tendency error, but no difference in halo effect. Both rating scales were significantly correlated with test performance scores and no significant differences between coefficients were found.

Comparison of black and white performance on the scales and tests showed that the research ratings evidenced less bias.

The Value of Multiple Criterion Measures in Validity Studies

Michael Rosenfeld and Richard F. Thornton, Educational Testing Service, Princeton, New Jersey

The advantages of using multiple criteria in criterion-related validity studies were described. Some proposed advantages were:

1. They increase coverage of the different aspects of the performance domain.
2. Researchers can evaluate the effectiveness of each measure.
3. Multiple criterion measures increased the likelihood of finding validity if it does exist.

Entry-level police officers concurrent and predictive studies were conducted. Four different criterion measures were used: work knowledge, self ratings, training grades, and supervisory performance ratings.

The results were:

1. Work knowledge tests appeared to be an appropriate criterion for black and white officers and were fair (consistent with selection test results and no adverse impact).
2. Self ratings were not usable. Some officers rated themselves outstanding on all 37 tasks, and other response patterns emerged which showed they were not taken seriously.
3. Supervisory ratings were found to be an appropriate criterion for white police officers only. Poor reliabilities and poor correlations with other performance measures were found when black officers were evaluated.
4. Training grades were significantly correlated with test selection test scores for both blacks and whites and were very similar values.

In conclusion, it was observed that if they had used only supervisory ratings they would have had a much more difficult time defending the validity of the test than with the multiple criterion measure approach.

The Effects of Test Validity on Workforce Productivity

Murray Mack, Information Science, Inc.
Frank L. Schmidt, U.S. Office of Personnel Management and George Washington University

The results of a study assessing the dollar impact of four different models of selection were reported and discussed: random selection from the total applicant pool, random selection above low minimum cutoffs, optimal use of tests, and quota top-down selection. Test validity was held constant across the four selection strategies.

Three of the four strategies (all but optimal test use) are used for affirmative action to aid in acquiring a representative work force. The dollar impact on work force productivity of these selection strategies was examined in the context of the job of park ranger. Part of this research required estimates of the standard deviation of job performance in dollars by first-line park ranger supervisors. Formulas for assessing dollar benefits of productivity were discussed and are presented in publications by Mack and Schmidt. Other utility studies have been conducted on the job of computer programmer and budget analyst.

According to the findings of the park ranger study, one of the three affirmative action strategies, quota top-down selection, appears to maximize minority hiring and at the same time minimize the loss in utility, while the other two affirmative action strategies are "economic disasters in terms of work force productivity."

Quota top-down selection leads to minority hiring at the same rate as majority hiring, with almost 94% of the gain in productivity associated with optimal test use. The reason top-down quota selection and the random selection above low cutoff models are so different is that the productivity losses are not due to the fact that fewer or greater minorities are selected but rather due to the randomness in selection.

ASSESSOR TRAINING: THE LEADERLESS GROUP DISCUSSION

Dennis A. Joiner
Dennis A. Joiner & Associates, Sacramento, California

Mr. Joiner is a consultant in assessment centers. He discussed his training approach and then presented a videotaped training exercise for assessors. His approach is a three-day model involving sending out comprehensive reading materials to assessors one week in advance and then one day of on-site training the day before the assessment (10 hours), one full day of assessment, and one day of integration, final evaluations, feedback, and final rank order listing. He evaluates 10-12 candidates per day of assessment and increases the number of assessment days depending on the number of candidates.

Assessor training includes:

1. Information about the job and how it fits in the agency. He links performance dimensions of the test to tasks.
2. A detailed review of examination materials.
3. Practice in recognition, observation, classification, and grading of behavior. The majority of the training is focused on this area.
4. The assessor's role, how all exercises fit together, even if an assessor only participates in one or two.

He also trains assessors to avoid typical rater errors such as:

Halo Effect: Differentiate between strong leadership and quality of decisions and analysis.

Projection Errors: Favoring one style or approach when several will do. Avoiding the one "best" way.

Impression Errors: Assessors should evaluate the total session not individual incidents.

Stereotyping: If they find themselves noticing things that are not relevant to the dimensions, they are asked to consciously avoid considering these things.

Contract Error: Errors and differences between candidates are to be avoided.

High, Low, and Central Tendency Errors: Avoid the "benefit of the doubt"; use the whole scale.

All candidates are rated by two assessors because of the difficulty of evaluation.

Given this information and additional written instructions, session participants then viewed Joiner's 33-minute assessor training videotape on leaderless group discussion.

MICRO COMPUTER DEMONSTRATION

Chair: Phil Carlin, City of Tucson

Demonstrators: Theodore S. Darany, San Bernardino County Personnel, California
Jack Feldhaus, Pima County Personnel Department, Tucson, Arizona
Stephen J. Mussio, Minneapolis Personnel Department

This session was enthusiastically received by the large number of attendees. Computer programs which were demonstrated were of particular interest because they were for use in the area of personnel assessment.

Ted Darany demonstrated programs for use on an Apple computer. Jack Feldhaus demonstrated programs for use on a Commodore computer. Steve Mussio demonstrated programs for use on an Ohio Scientific computer. A company representative demonstrated programs for use on a Radio Shack computer.

The overwhelming impact or idea which emanated from these demonstrations was the tremendous capability that can be obtained for very little cost. A person could easily make the observation that the term "micro" refers to the cost rather than the functional capability. Programs that were once in the domain of large centralized data centers can now be readily run on a desktop computer.

Readers of these proceedings are encouraged to contact the demonstrators for details on their particular software and hardware configuration.

Theodore S. Darany
Employment Division Chief
San Bernardino County Personnel
157 West Fifth Street
San Bernardino, CA 92415

Dr. Jack Feldhaus
Personnel Psychologist
Pima County Personnel Department
151 West Congress, 4th Floor
Tucson, AZ 85701

Stephen J. Mussio
Director, Evaluation Services
Minneapolis Personnel Department
312 Third Avenue South
Minneapolis, MN 55415

PAPER SESSIONS

PERSONNEL ADAPTATIONS OF VIDEO

Chair: Cassandra Scherer, Milwaukee Police and Fire Commission

Discussant: Dave Lookingbill, Nebraska

Standardizing Job Analysis Through Television

William Tomes, South Carolina Merit System

In this presentation, Mr. Tomes discussed two applications of television to the job analysis process and presented one of the videotaped products.

In one product, a state-owned television system was used to facilitate the generation of tasks and KSAs through a multiple-site brainstorming session with two-way communication to the television studio. Although this approach worked, it was felt that this process was not quite as effective as an on-site session. Time was spent by subject matter experts passing the microphone back and forth and relaying information to the television studio. Also, subject matter experts' voices were indistinguishable from each other. The overall cost savings, however, may make it an alternative to consider if the facilities are available.

In another project, television was used to conduct sessions where the SMEs completed their application questionnaires. In this project, the sessions were conducted using both live broadcasts and prerecorded videotapes. SMEs were mailed the questionnaire and were able to watch the same set of instructions everyone else was receiving. With the live broadcast, two-way communication and question answering was possible. This was not true of videotape, but both techniques seemed to work very well and resulted in considerable overall cost savings and improvement in the error reduction in filling out the questionnaire.

Video Testing for Entry-Level Hospital Attendants

Bob Schneider, Pennsylvania Civil Service Commission

In this presentation, Mr. Schneider discussed the development and administration of videotaped exams used to screen entry-level attendants in state-operated institutions. Task analysis had determined that applicants needed only eighth grade reading and writing skill, ability to be trained, and attitude for the job. Researchers decided to use a content validity approach and present elements of the actual training program to the job applicants as a work sample test on the theory that, if applicants can master the training materials presented in the test, they can master training on the job.

After addressing some of the problems encountered in using paper and pencil exams for entry-level attendants, Mr. Schneider indicated how a videotaped exam minimized or eliminated many of these problems. Reasons for using videotape were:

1. Accessible equipment and technical support
2. Ability to present visual example of work setting and environment to candidates, including positive and negative aspects of the job to encourage self selection
3. Ability to overcome lack of reading skills among candidates
4. It was theorized that since TV is a part of everyone's life, it would be less threatening than a written test. This aspect did not work well. Instead, applicants tended to forget they were in a testing environment, they became absorbed in the events of the test, and talked to each other conversationally.

Mr. Schneider also discussed some problems they were working on:

1. All applicants must proceed at the same pace. They have geared the time to the slowest candidate in the pretest.
2. A restricted number of TV monitors limits the number of candidates who can view the test at one time without more sophisticated TV equipment.
3. Technical problems with the equipment can be a problem in isolated exam sites.
4. Correction and modification of the test will be expensive and time-consuming.
5. It is not possible to test those who cannot hear or see; however, this is likely a job-relevant skill for this case.

The session concluded with excerpts from the actual exam tape being shown.

CONSIDERATIONS CONCERNING DISCRIMINATION

Chair: Sinclair Hugh, Denver

Discussant: Ernie Long, Office of Personnel Management. Seattle

Effective Goal Setting in Affirmative Action Programs

Joseph Ivers, Wayne State University, Detroit
Andres Inn, Advanced Research Resources Organization, Washington, D.C.

The need exists for a method of systematically setting affirmative action goals. A computer simulation method is proposed and three different affirmative action approaches are examined using the computer model. In the

model, organizational parameters are specified as well as operational definitions of various affirmative action policies. The mobility of employees within the organization of interest is then simulated across varying lengths of time to determine the most likely outcomes of pursuing a given policy. Optimal or less than optimal conditions of hiring or promotion along career paths can be simulated, subject to the assumptions one wishes to make about factors such as availability and length of tenure. The possible outcomes of various policies were then evaluated and compared.

Three hiring models were considered:

1. Merit hiring
2. One-to-one quota, where one protected group member is promoted for every nonprotected group member.
3. Two-to-one quota, where two protected group members are promoted for every one nonprotected member.

Observations were projected across twenty years in one organization and across five years among sixty organizations. The model organization was pyramidal with entry-level capacity of 120, middle level of 20, and top level of 5. Vacancies were the result of promotion and attrition.

Resulting analysis showed that few substantial trends appeared during the first five years, but that trends did develop in the middle and upper levels to increase representation of women and minorities after ten or more years. It was concluded that, given the slow turnover time implicit in small, closed systems, the effects of a consistently pursued EEO program will take considerable time to come to fruition. Affirmative action plans are indeed critically important. At present, it appears that all parties interested in the success of affirmative action plans stand to benefit from a systematic approach to policy evaluation and goal setting attainable by means of Markov-related simulations such as this.

Assessing Employment Discrimination in Noncompetitive Promotions

Chuck Martin, Goulding, Martin & Associates, Houston, Texas

Recent court decisions have placed promotional practices under increasing scrutiny. Analyses of discrimination in noncompetitive promotions raise new issues as to eligibility for and timing of promotions. A variety of conclusions can be drawn from the same set of data depending on how promotion elibility is defined.

Some issues in reclassification are: Why is a person reclassed? For example, is the job being reclassed or the person being reclassed? Is reclass the function of selection factors? In reclasses, all original selection data are ignored; however, the highest ranking at entry may be the best future performers as well. There may be room for use of Bayesian approaches here.

Reclassification is a function of the supervisor, the workload of the organization, and cyclical trends in the organization. Because of problems of N=1 in reclass situations, any analysis of discrimination in reclasses would have to be based on combined data from dissimilar jobs. This is well known as a poor statistical approach.

In one example Martin cited, trend analysis was applied to show why minorities spent more time in grade before reclassification than whites did. In this case, blacks were available and hired in June or July, but workload occurred in January. More whites were hired later in the year.

There are simple Bayesian models that will allow you to combine factors to analyze reclassifications. Martin encouraged increased application of more sophisticated statistical analysis to those problems.

Effects of Race and Sex Role Stereotypes Upon Intra-Interview Assessment

Charles Ridley, Personnel Decisions, Inc., Minneapolis

Defensive reactions of black interviewees or female interviewees may occur as a response to racist sexist behaviors and stereotyped attitudes of an interviewer. In turn, the interviewer may misinterpret these reactions on the part of the client as pathological behavior. More critical implications are noted:

1. Theory explication: The interpersonal behavior of women and black Americans cannot be presumed to exist in a psychological vacuum. Behavior analyses based on interviews demand sensitivity to racism, sexism, and other environmental factors which negatively impact on interviewees.
2. Research methodology: A call for a new model of research is in order. Most of the research in this area has been concerned with either therapeutic process or therapeutic outcome but not both. Methodology is needed which combines process and outcome variables into single investigations of the interviewer-interviewee interaction.
3. Assessment models: Interviewees no longer can assume that employment decision making is totally an objective process or that their training has immuned them from racist and sexist biases or idiosyncrasies. The author proposes the development of a diagnostic classification system which is sensitive both to specific reference group traits and social determinants of behavior. This new race and sex sensitive taxonomy should provide a more reliable and valid system for assessing the behavior of black and women interviewees.
4. Graduate training: In recent years, the need for graduate training to recognize sexual and racial variability has been pronounced (APA Task Force on Accreditation, 1979). Interviewers in training should undergo personal counseling with a therapist of the other race or sex who is skillful in race or sex psychology. The purpose of this endeavor would be to identify, understand, and eliminate sexual/racial stereotypes.

Comments of Discussant

Ernie Long, Office of Personnel

Based on the results of the first paper, affirmative action progress will have to come from outside hires, not promotion within.

Reclass issues in court cases seem to be that the same criteria were used in all cases. This requires that they be clearly documented and applied. This would increase the likelihood of defensibility in court.

If interview content and protocols are valid, he would be surprised if the effects mentioned in the Ridley paper would affect the outcome significantly.

The push for validation of interviews has resulted in highly job specific selection criteria. Could they be too specific? Exposure to situations within the organization can give the benefit to current employees of the organization who are more familiar with the specifics than nonemployees. What is the impact of this on affirmative action? He prefers that candidates report past behaviors which show each trait, rather than respond to job-related situations.

OPEN FORUM

Five-minute talks on any topic relevant to IPMAAC

Open to all conference attendees, as time permitted.
Speakers were requested to register in advance.

Chair: Lance Seberhagen, Seberhagen & Associates, Inc.

Resource Panel: Glenn McClung, Denver
Nancy Abrams, Office of Personnel Management, New York
and others

The open forum proved to be an excellent opportunity for conference attendees to ask questions and give input to the incoming IPMAAC board and committee chairs on the directions they would like to see IPMAAC pursue in the future. A large number attended the one and one-half hour session, and many excellent ideas were proposed for the next conference program, newsletters, and projects for exchange of information. A sampling of ideas presented includes conference sessions and articles on state-of-the-art research reviews, sessions that tie in selection with other areas of personnel such as comparable worth, providing a display of answer sheet scanners and test publishers in Minneapolis next year, and exchanging through newsletter or other methods examples of tangible selection products such as rating scales or instructions to oral boards.

SYMPOSIUM

The Development of Job-Related Medical Standards

Presenters: Deborah Gebhardt, Advanced Research Resources Organization, Washington, D.C.
David C. Myers, ARRO
John Kohls, California Commission on Peace Officer Standards

Discussant: Charles Dotson, University of Maryland

According to Deborah Gebhardt from Advanced Research Resources Organization (ARRO), the increased emphasis placed on hiring qualified personnel for physically demanding jobs has created the need for a personnel selection process which evaluates the physical capabilities and tells the status of an applicant in relation to the job requirements. One facet of this selection procedure is the pre-employment medical examination given by a family or company physician.

The purpose of this research was to supply the physician with physical performance test scores and an avenue for integrating these test scores. The physical performance tests that were developed were systematically linked to biomechanical and physiological demands associated with the jobs under review. The job performance standards for these tests were established and incorporated into a manual. This physician's manual included a brief outline of the job and its physical demands, the physical ability tests and standards, and the literature review of the current biomechanical and physiological research related to the demands of the job.

The development of this physician's manual provided the physician with a more objective procedure for identifying an individual's physical capabilities. Additionally, this research assisted the physician in making valid pre-employment judgments by providing a content valid framework based on the physical demands of the job.

David Myers, also from ARRO, summarized an approach for linking the type and degree of medical impairment with the physical requirements of jobs in order to develop guidelines for medical examiners. The goal of this approach was to provide an empirical and rational basis for relating physical demands of jobs to specific abilities and, in turn, relating those physical abilities to medical conditions. From this perspective, a direct connection can be made between specific medical conditions and qualification or disqualification of job applicants.

The first step of this process was the delineation of fourteen basic physical abilities (e.g., dynamic strength, flexibility, and equilibrium) which can be required in job situations and rated on a seven-point scale to measure the degree or amount of each ability used. Second, 96 typical medical diseases representing nine major body systems were chosen. The relationship between the medical conditions and the physical abilities

were defined in two ways: the limitation of the worker's physical ability to perform physically demanding tasks adequately and the hazard to the worker in possibly aggravating the disease while performing the tasks.

Next, medical experts (15 students in their final year of medical school) were given a manual listing the physical abilities and the diseases. For each of the physical abilities, a grid was presented with diseases listed horizontally and the seven-point scale of degree of physical ability required listed vertically. The medical students were asked to select the cutoff point on the scale to indicate at what level of task performance a person with the particular disease would be able to perform all less demanding duties and would be able to perform all more demanding duties. They did this for all medical diseases and all physical abilities, and the results were averaged. There was substantial interrater agreement with static strength having the highest reliability (.92) and equilibrium the lowest (.77).

With the information from the process, it is possible to analyze a particular job and determine the type and degree of its physical requirements and which physical impairments would significantly hamper its performance. Although the 96 impairments used do not exhaust all possible diseases and injuries, they can act as a benchmark for examining physicians who might not otherwise have an adequate appreciation for the requirements of the job. The need was also pointed out for better definitions of diseases and levels of severity with the suggestion that the American Medical Association's "Guide to the Evaluation of Permanent Impairment" might be a useful classification tool.

John Kohls discussed the efforts by the California Commission on Peace Officer Standards and Training to develop job-related medical standards using the expert judgment of physicians and the documentation of the physically demanding aspects of the job as well as considering relevant legal and fair employment practices. He highlighted the following six legally defensible reasons by which an applicant could be rejected. These were: (1) if the condition (i.e., medical/physical) results in the inability of the applicant to perform the physically demanding aspects of the job; (2) if there would be a heavy burden to accommodate the job to the applicant's disability; (3) safety and risk factors affecting the applicant or others in the employment setting; (4) environmental hazards which would be dangerous to the applicants' health because of the given disability; (5) the probability of time loss due to the condition in excess of allowable sick days; and (6) the increased probability of early disability. It was noted that these last two areas are much more difficult to assess and accordingly more difficult to legally defend.

The discussant, Charles Dotson, addressed the fact that it is difficult to separate whether physical or medical indicators are causing the greatest problems in job-related settings. Also discussed was that further research should be conducted on the people who are screened out by the medical process.

SYMPOSIUM

Survey of Basic Item Analysis

Moderator: Nancy E. Abrams, U.S. Office of Personnel Management New York

Presenters: Leroy B. Sheibley, Pennsylvania State Civil Service Commission
Theodore S. Darany, San Bernardino County, California
Jerry J. Durovic, New York State Department of Civil Service

The first presenter, Leroy Sheibley of the State of Pennsylvania. discussed the "hows and whys" of item analysis based on the Pennsylvania "Examiner's Manual" Chapter VIII, Section B. He compared the test developer to a detective using the item analysis to provide clues about a test. He discussed the Pennsylvania item analysis which includes frequency by alternative for high and low groups (using a median split), item difficulty (p), and subtest and total test point biserial correlations. The test analysis includes number of items at each difficulty and item/total test correlation level, number of items at each choice position, average difficulty by choice position, mean, standard deviation, and KR-20 reliability. He discussed the use of these statistics in decisions to reconsider key, to change, to delete, or to keep items. Finally, he discussed the use of analysis by race, sex, and educational level.

The second presenter, Ted Darany from San Bernardino County, California, discussed the use of item analysis in a smaller jurisdiction. Many of the statistics discussed were similar to those of the first presenter. He did discuss the reasons why looking at the top and bottom 27%, rather than a median split, is preferable. In a small jurisdiction, item statistics often must be aggregated until a sufficient number is obtained, usually a minimum of 50. The presenter stated that no one regardless of size of jurisdiction should be developing tests without the use of basic item analysis data.

The final presenter, Jerry Durovic from New York State, presented examples of item analyses used in his jurisdiction. Similar statistics to those of the first two presenters were discussed. He noted that New York State conducts analyses by subtests only since each subtest should be measuring different content areas. Additional statistics, including IRI (item reliability) and coefficient alpha were discussed. The presenter discussed how analyses by race, sex, and ethnicity often point up problems in interpretation of item content. He emphasized that test analysis should be viewed in terms of the intended results of the test, that is, whether the test was intended to differentiate those at the top, middle, or bottom end of the distribution.

Each presenter volunteered copies of examples of his jurisdiction's item analysis for those interested.

INVITED ADDRESS

Sponsored by the Western Region Intergovernmental Personnel Assessment Council

Host: Roger Carey, Personnel Department, Solano County, California

Truth-in-Testing Legislation

Rexford Brown
Educational Committee of the States, Denver

Dr. Brown traced the history of truth-in-testing legislation from the first law enacted in 1978 in California, which applies to standardized tests used for postsecondary education admissions selection and requires disclosure of information about test features, test limitations, test uses, etc., to the second law enacted in 1979 in New York, which applies to tests used for postsecondary or professional schools admissions and requires full disclosure of the background information on the test, test items, test answers, etc., to the current pending federal legislation sponsored by Representative Gibbons (which applies to postsecondary admissions and occupational testing, does not require full disclosure, and does not allow norm-referencing) and Representative Weiss (which applies to standardized postsecondary admissions and placement tests and requires full disclosure). He stated that 24 states are also considering truth-in-testing legislation. According to Dr. Brown, the debate about testing falls into four major clusters of arguments, each with highly vocal proponents and opponents with strong, polarized positions. The four clusters and some of the arguments are:

1. The Role and Power of Testing Companies

Pro-Testing Arguments	Anti-Testing Arguments
Tests have little influence compared with family, social and educational influences, GPA.	Tests have profound influence upon American lives and life chances
Commercial test publishers are accountable to market forces; test makers, including ETS and ACT, are accountable to professional standards, education community, higher education communities, courts, client groups, trustees, and IRS.	Testing companies are unaccountable to their dependent public--particularly true of ETS and ACT.
The public and higher education have asked for massive testing; testing produces information useful for improving education; it does more good than harm (takes little time, diagnoses problems, helps administrators, etc.).	Massive testing does more harm than good (consumes time better spent learning, alters curriculum, stigmatizes children, misleads public, etc.)

Pro-Testing	Anti-Testing
Test companies try hard to curb abuses, educate users.	Tests are widely misued and misunderstood.

2. Quality of Standardized Machine-Scored Tests Used Primarily for Prediction

Defenders of Standardized Tests	Critics of Standardized Tests
Society values intellectual achievement cognitive skills; education (especially higher education) stressed those skills; others (e.g., teachers) are better able to assess imagination, creativity, etc.	Tests concentrate on easily measured cognitive skills, ignoring higher level skills (problem solving, imagination, creativity, etc.).
Theory upon which education rests may be simplistic, outdated, and sketchy; test theory is better than critics think and always improving.	Theory upon which testing rests is simplistic, outdated, and sketchy.
Many tests are rigorously validated and most do what they are designed to do.	Tests are seldom valid even by test makers' standards.
Tests are developed by educators and scholars, some of whom always disagree with others; in the main, they do what they are supposed to do.	Tests are developed subjectively and always contain controversial items.

3. The Need for Testing Legislation

Pro-Legislation Sentiments	Anti-Legislation Sentiments
A commitment to "truth in lending," "truth in advertising," sunshine laws and consumerism should extend to an area as important as admissions testing.	Test publishers and higher education institutions already provide ample information and protection; analogies to consumer movements are misleading.
Because admissions tests have such influence, there is an overriding public interest at stake.	There are several competing public interests at stake; critics have not established an overriding need for legislation.
Legislation will promote greater accuracy and validity of tests.	Legislation calling for full disclosure will lower the quality of tests.
Legislation will encourage use of multiple criteria in selection process.	Most institutions already use multiple criteria and test agencies encourage the practice.

4. Full Disclosure and Other Aspects of Legislation

Arguments for Full Disclosure	Arguments Against Full Disclosure
Students can learn about tests and test strategy from examining test questions.	Students cannot learn much from examining their test items.
Security need not be an issue; new measurement technology could enable testers to eliminate the problem.	Full disclosure will compromise test security; compromised security means less confidence in tests.
Development costs would not increase as much as testers suggest.	Disclosure will increase test development costs, thus the cost of tests to students; poorer students will suffer.
The fairness issue takes precedence over technical matters.	Disclosure requirement constitutes seizure to private property without due process and in violation of proprietary rights protected by copyright laws.

Dr. Brown then discussed some questions about testing and its role in our society: Do admissions tests seriously affect the life-chances of American students? Are admissions tests being misused/misunderstood/misadvertised in ways that place some students at a disadvantage? Will regulatory legislation requiring full disclosure soften the influence of tests upon life-chances and help correct problems stemming from misuse? Depending upon one's perspective of the issue involved, the answer to each of these questions is yes, no, or maybe.

Dr. Brown concluded his remarks by suggesting that certain questions must be answered, with empirical evidence where possible, before the need for regulatory legislation can be decided:

1. Is there a "problem" with testing that requires state or federal legislative action?
2. Is open testing technically feasible? for all tests or only certain ones?
3. What will be the consequences of open testing for various kinds of tests and levels of education?
4. Will open testing erode or enhance confidence in the tests? with what result?
5. Will any personal, social or educational benefits of open testing be offset by such problems as decreased validity, increased cost or reduced use of the tests?
6. Do test companies already release sufficient information?
7. Are there partial disclosure plans that would be preferable to full disclosure?

8. What is the magnitude of standardized test misuse? in admission to postsecondary schools? professional schools? elementary and secondary education?
9. Will disclosure legislation correct misuse?
10. What legal and constitutional problems are raised by which provisions in current testing laws?

It is clear that both proponents and opponents of legislation have lots of homework to do. In Dr. Brown's opinion, if the legislative debate does nothing else than bring more information and clarity to this public policy area, both sides will have been well served by it.

(Note: Dr. Brown's report, "Searching for the Truth About "Truth in Testing" Legislation," is available for $6.50 from the Education Commission of the States, 1860 Lincoln Street, Suite 300, Denver, CO 80295.)

INVITED ADDRESS

Sponsored by the Great Lakes Assessment Council

Host: Sally MacAtee, Personnel Department, City of Milwaukee

Selection Research as Seen by a Personnel Director

Chuck Grapentine
Administrator of Personnel, Wisconsin Bureau of Personnel

Mr. Grapentine began his remarks by stating that there is not so much a lack of selection research as "an overabundance of junk" which is serving little or no administrative purpose. In the meantime, other needs are being neglected. The research that has been produced is insufficient to use in defense of complaints of discrimination, insufficient to persuade labor negotiators, and insufficient to reduce adverse impact.

The question receiving the most attention in this research has been how to meet the requirement in Section 3B of the Uniform Guidelines on Employee Selection Procedures which demands that an employer must use the selection procedure with the least adverse impact given equal validity. Mr. Grapentine traced the history of this research from the investigations of differential validity through attempts to reduce test anxiety and to increase test sophistication to the present when ". . . we still find enforcement agencies clinging (for) dear life to the worn-out idea that the problem is the test instrument . . . ." The unhappy fact is that many professionals are still willing to expend resources on such investigations.

Mr. Grapentine stated that continuing to seek a research-based solution to the question of adverse impact is ill-advised because (1) there is little hope of success as demonstrated by past and present data; (2) other demands must be addressed by research and administration; and (3) there is an administrative alternative. He proposed expanded certification (adding the names of the three highest scoring minorities or women to the normal cert list for each vacancy in a class that is not racially/sexually balanced) as a solution. Realizing that management's responsibilities include conflict resolution and balancing resources, demands and needs, and the need to simultaneously pursue affirmative action and merit hiring, he felt this suggestion deserved a try.