DOCUMENT RESUME

ED 337 479 TM 017 293

TITLE Proceedings of the 1982 IPMAAC Conference on Public Personnel Assessment (6th, Minneapolis, Minnesota, June 6-10, 1982).

INSTITUTION International Personnel Management Association, Washington, DC.

PUB DATE Jun 82

NOTE 93p.

PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC04 Plus Postage.

DESCRIPTORS *Evaluation Methods; *Job Performance; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; Predictive Measurement; Private Sector; Psychological Testing; *Public Sector; 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. Author-generated summaries/outlines of papers presented at the IPMAAC's 1982 conference are provided. The presidential address is "Professionalism and Productivity" by G. G. McClung. The keynote address is "The Validity of Content Valid Tests and the Basis for Ranking" by J. J. Hunter. The following paper sessions are summarized: "Automation in the Personnel Office"; "Technological/Methodological Issues in Selection"; "Selection Procedures in Police and Fire Settings"; "Performance Appraisal and Evaluation"; "Innovation and Alternative Selection Procedures"; and "New Developments in Selections". The following symposia are summarized: "Labor Unions: How Much of an Effect on Personnel Assessment Decisions?"; "Development of Job Related Minimum Qualification Requirements--Methodological and Administrative Concerns"; "Validation of Selection Procedures with a Small N: An Example of a Successful Approach"; "Ethical Considerations in Personnel"; "A Legal and Technical Analysis of Issues Involving the Ranking of Candidates and Setting of Passing Points"; "Productivity, Effectiveness, and Performance Appraisals: Commonality or Confusion?"; "The Use of Written Simulations in Personnel Selection"; "Professional Accountability to Applicants"; and "Psychological Testing: Its Survival Problems". An invited luncheon speaker's paper is summarized: "Public and Private Sector Assessment: Is There a Difference?" by V. R. Boehm. The Western Region Intergovernmental Personnel Assessment Council's paper "The Cutting Edge of Selection Developments by R. L. Ebel; "A Bayesian Approach to Validity Generalization: A Systematic Examination of the Robustness of Procedures and Conclusions" by K. Pearlman; and the Great Lakes Assessment Council Panel on "Selection in the Private Sector" are reviewed. Summaries of two full-day and two half-day preconference workshops are also included. (SLD)

IPMA Assessment Council

IPMA ASSESSMENT COUNCIL

PROCEEDINGS OF THE

1982 IPMAAC CONFERENCE

ON

PUBLIC PERSONNEL ASSESSMENT

JUNE 6-10, 1982

MINNEAPOLIS, MINNESOTA

PROCEEDINGS OF THE 1982 IPMA ASSESSMENT COUNCIL CONFERENCE ON PUBLIC PERSONNEL MANAGEMENT

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

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

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

PREPARED UNDER THE GENERAL DIRECTION OF:

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

Credit for major assistance in the summarizing of reports, and other aspects of the compilation of the proceedings, goes to:

Nancy E. Abrams, Consultant, New York
Michele Fraser, Personnel Decisions, Inc., Minneapolis, Minnesota
Telma Hunt, George Washington University
Lynn Robbins, Psychology student, George Washington University
Barbara A. Showers, Director, Office of Examinations, State of Wisconsin

IPMAAC EXECUTIVE COMMITTEE
Charles B. Schultz, President
Barbara A. Showers, President-Elect
Glenn G. McClung, Past President

IPMA ASSESSMENT COUNCIL

The INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL (IPMAAC) is a professional section of the International Personnel Management Association--United States for individuals actively engaged in or contributing to professional level public personnel assessment.

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

- providing opportunities for professional development;
- defining appropriate assessment standards and methodology;
- increasing the involvement of assessment specialists in determining professional standards and practices;
- improving practices to assure equal employment opportunity and merit employment;
- assisting with the many legal challenges confronting assessment professionals; and
- coordinating assessment improvement efforts.

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

TABLE OF CONTENTS

Page

PRESIDENTIAL ADDRESS

Professionalism and Productivity

Glenn G. McClung, Denver Career Service Authority

According to IPMA and IPMAAC literature, we are an organization of personnel "professionals." This is both convenient and provoking. It is convenient because it allows us to confer the name "professional" on anyone willing to join and pay dues. It is provoking for the same reason.

President McClung indicated that as his term as IPMAAC President comes to a close, he finds himself introspective on the subject of "professionalism." What does it really mean, who qualifies, and how does it interact with the demand for productivity in the workplace?

The term "professional" is often used to describe an individual in any occupation who displays the highest standards of commitment, competence, responsibility, and concern for others. For example, it is said that in the theater, a professional is "one who goes on stage and does his job, whether or not he feels like it." McClung stated that he was quite fond of this definition, but that it is obviously very broad and not very exclusive on an occupational basis.

In 1912 Supreme Court Justice Louis Brandeis offered a set of three criteria for defining a professional occupation. First, the necessary preliminary training is intellectual in character. Second, a professional occupation is pursued largely for the sake of others. Third, the accepted standard of success is not financial success. In the 1960's, in an article entitled "Is Personnel Administration a Profession?", Dr. Thomas Patten added two more criteria to those set forth by Justice Brandeis. Dr. Patten suggested that any profession should have an association of members whose chief functions are the enforcement of professional standards of behavior and the advancement and dissemination of knowledge. A profession should also proscribe certain minimum standards of training and competence. President McClung pointed out that personnel professionals need to establish both minimum requirements for entrance into the field, and a code of professional conduct.

President McClung acknowledged the dangers inherent in establishing minimum standards of training and competence. Some professions have used such standards to close their ranks in an exclusionary fashion for the supposed benefit of the public. Those in the personnel field have tried to eliminate artificial barriers to employment to avoid such self-serving credentialism. In establishing minimum standards it is necessary to achieve a balance between efficiency and productivity on one hand, and fairness and egalitarianism on the other. In personnel assessment it is a daily task to achieve this balance as passing points and minimum entrance requirements are established for specific classes of work. If standards are lowered in the interest of fairness, a greater number of unproductive people are likely to be employed. If standards are raised, a greater number of people who are competent are likely to be rejected for the job. McClung asked, "In our zeal for the elimination of credentialism and artificial barriers, how many valid standards may we also have abolished... and what may we have done to our own field?"

Recently it has been popular to allow liberal substitution of experience for education, thus, encouraging promotion from within. This has improved opportunities for employees at lower ranks and avoided the costs of open recruitment and screening, but it has also excluded well-trained people in search of an entry-level professional job.

A major difficulty in establishing a professional code of ethics is that personnel specialists serve a wide variety of clients. For example, those employed by independent merit-system agencies have at least three sets of clients. The first is the political leadership of the jurisdiction, the second is the employees of that jurisdiction, and the third is the voting public. It will be extremely difficult to develop a code of professional behavior that can cover these varied circumstances. Yet, a general set of principles is needed to avoid opportunism and situational ethics

Mr. McClung concluded his remarks by pointing out that people join professional associations for many different reasons. They get their names placed in the directory, they get out of town once in a while, and they try to have fun. Most of all they join because they care about what they are doing with their worklife, see the implications beyond their paychecks, and want to improve their practice. This is what professionalism and productivity are all about!

PAPER SESSION

Automation in the Personnel Office

Chair: Fay Walther, City of Forth Worth
Discussant: Jennifer French, San Bernardino County

Two Cost Saving Computer Applications in Personnel: Word Processing and Data Processing

Donald A. Emmerich, City of Dallas Civil Service

Microprocessor systems have much to offer the field of personnel assessment. Mr. Emmerich illustrated many of these benefits with the test construction procedure used by the Dallas Civil Service. Use of word processors in this procedure have saved time, reduced the number of errors, and provided for greater security of test materials. Other advantages include multiple workstations allowing simultaneous use of the system for different purposes, ease of operation, minimal "down time," and no information loss during three years of use.

Word processing systems often do not have data processing capability, but these functions can be obtained through the addition of more hardware. Having these capabilities in the same system provides for more comprehensive functions. For example, Mr. Emmerich pointed out that if test items are stored on a word processor, it is useful to have item statistics stored in the same location. If the system has data processing capabilities, the item results can be computed and posted by the machine. Different functions that are of use in selection and examination include item analysis and tabulation, score analysis, test grading, and administrative functions.

Mr. Emmerich also discussed the advantages of microprocessor systems over the use of mainframe computers. Software is available for purchase or can be easily developed for microprocessor systems, while development of software for a mainframe can be time consuming and expensive. Mainframes are subject to "down time," and lose information on occasion. Microprocessor systems utilized by personnel departments can be programmed and operated by the personnel staff to fit their specific needs, while mainframes are usually programmed and operated by computer specialists who lack a thorough understanding of the needs of the personnel staff.

The cost effectiveness of microprocessor based data processing systems makes this approach very attractive for many situations. One example is the small to medium scale applicant test paper processing activity.

Many of the basic operations, including test paper scoring, item tabulation and statistics calculation, test score analysis (distributions, pass-fail rates, ethnicity, and sex analyses), and so forth, require eight to ten minutes per applicant. Inexpensive equipment is available which can perform these operations in approximately ten seconds per paper; and, by inexpensive, Mr. Emmerich means less than $5,000 for hardware cost. If an agency examines 10,000 applicants per year, this examination requires approximately 42 work-weeks for these functions; microprocessor based equipment can perform these operations in approximately 28 work-hours. If the personnel cost in terms of salary and benefits to perform these operations is $15,000 annually, the microprocessor based system can pay for itself in approximately five months. This payback time to the agency does not take into account increased productivity resulting from other tasks that the test scoring individual is now free to perform since he or she has now relegated them to a machine. These estimates are approximate, and Mr. Emmerich believes that they are conservative. These figures also reflect using the equipment solely for test scoring and analysis functions. In reality, the equipment is far more flexible and can accomplish many other tasks.

Task Inventories and Task Taxonomies: A Problem for Job Analysis and Job Classification

Doug T. Goodgame, Texas A&M University

In 1961 The U.S. Air Force completed three years of research and concluded that the task inventory approach for constructing job analyses would produce valid and reliable job information. The volume of data generated by such an inventory can be quite extensive, and requires the use of high speed and high capability computers. The Air Force developed a series of computer software known as CODAP to process task inventory data. Due to changes in job analysis and computer technology, the CODAP system is being improved. A series called CODAP80 is being developed to handle some problems in job analysis which were beyond the capabilities of the original system.

One of the problems in job analysis is how to create a classification of tasks for each task inventory. Classifications can be developed for specific purposes and are organized around predefined characteristics or traits of the tasks. Unfortunately, analysts do not often have clearly defined characteristics for classifying tasks. The principles for classifying tasks will vary from one occupational field to another, depending on the definition and use of job specialties.

Tasks which are highly correlative normally indicate areas of work specialties unique to the occupational field. In essence, the duty fields of a task inventory represent a guess by the task inventory developer as to how the tasks will cluster together. A job clustering routine in CODAP80 brings together clusters of workers performing similar sets of tasks and is one of the first steps in a study to evaluate job class structure.

Difficulties arise when workers, who cluster together and represent a set of work specialties, check and rate [illegible] in duty fields which poorly reflect how their tasks are organized by these work specialties. CODAP80 can produce job descriptions at the duty level which reflect sums of time-spent values for tasks within each duty field. Duty level descriptions should help analysts quickly identify the work specialty of a cluster of workers on the diagram of job clustering actions.

One way to improve the classification of tasks is to identify tasks that are highly correlative. Factor analysis is too time consuming and expensive for this purpose. The alternative is cluster analysis. The characteristics of data matrixes in job clustering are well known. There has been little work done in the area of task clustering, but it is believed that cluster analysis can identify tasks that are highly correlative and lead to the identification of groups of tasks that "hand together" across various groups of workers. CODAP80 contains various routines for clustering occupational data, and Mr. Goodgame believes that those routines will prove useful in conducting the research necessary to identify the most useful routine for clustering tasks.

The Development of a Computerized Human Resource Management Information System

Reginald A. H. Goodfellow, California State University, Sacramento
Russell Nielsen, Riverside County Office of Education, California

The purpose of such a computerized system is to simplify the data requirements of managing human resources within an organization. This system was developed in the educational system in Riverside County, California, an educational system which employs a total of 30,000-32,000 individuals. (A regional data processing center was already in existence, which facilitated the development of the computerized system.)

The system has several components. The position control file contains all he information relating to various positions within the organization. The pre-employment file contains assessment and selection information pertaining to applicants for positions. Once an individual is selected and enters the system, he or she is included in the employee file. The post-employment file contains information on re-employment rights, retirement benefits, updating of credentials, etc. In each of these files there is detailed numerical coding of pieces of information.

This is run on an interactive data base system which is the crucial and critical aspect of the system. It is currently run on side-by-side "burro" systems which provide for 120 on-line, interactive, remote terminals. This system can be conceptualized as a series of bins, each containing different types of information. There is a position control bin, a pre-employment bin, an employee bin, and a post-employment bin.

The pieces of data can be accessed, taken out, used or modified, and put back again by anyone throughout the system. Because of the opportunity for data to run amuck or disappear, a full-time data base manager is employed to ensure the integrity of data entering and leaving the system.

The information in a data base system must be entered at the start, including all of the various options which could conceivably be necessary. It is difficult to enter additional information later. A position control data collection input form is completed for each position within an agency. Any and all information imaginable can be included, such as title, type, sequence, designation, status, EEO Five and Six designations, labor market data, budget codes, salary, schedules, benchmark status, and benchmark classification for comparability. There are equal amounts of information for each individual in the system. The two are tied together when a position code is attached to an individual. There is one person for each position.

The great advantage of this system is that all of this various information is immediately retrievable. An example of one practical use of this system is this: along with their final paychecks at the end of the year, each teacher in the educational system receives a final printout of his or her total annual cost to the organization in terms of both salary and benefits. This information can be computed quickly and easily on the data base system, while in most organizations it would take a tremendous amount of time and energy to compile this type of information.

This system helps in managing the human resources of the organization. For example, there are about twelve different exit codes for the various reasons that an individual might leave the organization. While sitting at a console, the user can call up this information by department and find out what happened to the last fifty people who exited and determined the frequency of exit codes. The user could also call up this information for a certain supervisor to see what pattern existed among his/her subordinates, or the user could look at a specific position across the whole organization. This system can be a valuable tool in areas as varied as budgeting, collective bargaining, EEO statistics, labor market information, employee comparisons, termination analysis, and personnel planning.

Goodfellow offered the following advice for anyone involved in developing such a system. Plan ahead. Establish small committees which stay constant throughout the project. Have patience. Remember that the system does not have to be large; a small organization can set up an effective data base system on a mini-computer. (A micro, however, is too small to adequately handle such a system).

In responding to a question concerning the security of the information, Goodfellow explained that in this system, there are ten levels of security. Each is only given access to certain areas of the data base.

There are also automatic checks built into the system to prevent, for example, a certain employee from giving him/herself a 100% salary increase. When certain parameters are exceeded, the information is double-checked automatically.

Goodfellow stated that the use and effectiveness of such a computerized human resources management system depends upon the sophistication of the user. Therefore, the training component is essential and should be stressed.

SYMPOSIUM

Labor Unions: How Much of an Effect on Personnel Assessment Decisions?

Chair: Nancy E. Abrams, Consultant, New York

Presenters: Jack R. Lawton, Wisconsin Department of Industry
Lance W. Seberhagen, Seberhagen and Associates
Peter Benner, Minnesota State Employees Union

The Involvement of Public Sector Unions in Selection: The Wisconsin Experience

Jack R. Lawton, Wisconsin Department of Industry

This paper represents the joint efforts of Jack R. Lawton and George W. Dawes, both working in the area of labor and human relations.

The purpose of this paper was to examine the current and future trends in the relationship between personnel staffing functions and activities and public sector bargaining. The authors reviewed the complexities and difficulties inherent in this relationship in the public sector within the Wisconsin experience. They described the current status of collective bargaining laws for both municipal and state employees, the development of these laws, the provisions which are related to employee selection, and current administrative and arbitration activity related to selection. For the purposes of their presentation, the term selection was defined in its broadest sense to include original appointments, promotions, transfers, demotions, layoffs, and recalls.

There is usually a tension associated with the relationship between merit system requirements and the public employee union's reliance on seniority as a basis for personnel decision-making. This is especially true for activities such as original appointments and promotions, where the merit system requirement is for selection based on objective, job-related criteria, while labor would prefer that those decisions be based on seniority.

The history of the development of collective bargaining rights for public employees in Wisconsin and other jurisdictions shows that unions have had a continuing interest in bargaining over staffing activities. Private sector unions have successfully implemented collective bargaining arrangements concerning such things as promotions and job evaluations. Recently, the distinctions have begun to disappear between the private sector and public sector.

As these distinctions disappear, public sector employee unions urge the adoption of private sector bargaining rights related to staffing activities. Since staffing decisions in any organization are inherently tied to the employee's pay, morale, and job security, it is not surprising that organized labor is vitally interested in how those decisions are made.

The presentation presented the detailed development of and provisions in public sector bargaining laws in Wisconsin to illustrate some of the public sector problems. Two general statements on the history were included in this summary.

In Wisconsin there are separate collective bargaining provisions for municipal and state employees. Wisconsin was one of the first states to enact a collective bargaining law for municipal employees in 1959 with the passage of the Municipal Employment Relations Act (MERA). MERA granted organizational representation and bargaining rights to municipal employees, and provided for bargaining on questions of wages, hours, and conditions of employment. However, both the courts and the Wisconsin Employment Relations Board made rulings that the initial legislation did not impose a duty to bargain upon municipal employers, enforceable through remedial orders. The development of municipal employee collective bargaining rights then proceeded through a series of amendments to the original act. Significant changes included a "no strike" provision, fact finding and mediation provisions, and the creation of an administrative agency (Wisconsin Employment Relations Commission) with the authority to enforce the duty to bargain in an expanded scope of bargaining. Changes enacted in 1977 included final and binding compulsory interest arbitration which was mandated for nearly all municipal employees.

The initial collective bargaining legislation for state employees in Wisconsin was enacted in 1966. This law included a limited scope of bargaining which specifically excluded compensation, fringe benefits, rules relating to promotions, layoffs and classification, and discipline from the scope of bargaining. It also included a comprehensive management rights clause typical of early collective bargaining laws. However, both management and labor found that the provisions of this initial legislation were too restrictive to allow for a meaningful bargaining and resolution of important labor issues; and, as a result, comprehensive amendments to the State Employment Relations Act were enacted in 1971. The changes included an expansion of the scope of bargaining to include wages and fringe benefits, and disciplinary grievances. The management rights clause became a permissive subject of bargaining; but, the scope of bargaining did not include aspects of merit systems related to initial appointments, promotions, and job evaluations. Because of the language which specifically restricts the scope of bargaining, the collective bargaining agreements, which the state has entered into with state employees, have not ventured into the full range of staffing activities that municipal employee agreements have.

Although the authors described several contracts that included merit-like provisions for promotions and transfers, there is little doubt that the tension between merit system selection and the collective bargaining agreements will persist. In fact, it will be heightened in at least the near future by the tight fiscal policies which will narrow bargaining activity to non-wage items currently defined as management rights. The situation is not improved by the lack of knowledge or appreciation most collective bargaining specialists have of selection which could lead to the accidental inclusion of testing methods or policy within union contracts. The worst place where selection methods could be reviewed would be in arbitration.

One possible way to avoid the hazards of bargaining for selection rights is to make sure that the union membership is not excluded from the exam development process. For this reason, the choice of subject matter experts is critical in test development for represented jobs. If the membership sees itself as having a noncontractual role in reviewing or writing questions, the management versus union issue is less likely to be raised in the selection process.

An I/O Psychologist's Concerns About Labor Unions and Employee Selection

Lance W. Seberhagen, Seberhagen and Associates

Ever since the days of Frederick Taylor, Industrial/Organizational (I/O) psychologists have generally been viewed as pro-management, and even anti-worker. The most extreme critics have gone so far as accusing I/O psychologists of exploiting the worker to gain maximum efficiency, productivity, and profitability. While it is true that most I/O psychologists work for management and do seek to increase efficiency, productivity, and profitability as a major part of their jobs, Seberhagen does not agree that most I/O psychologists are necessarily anti-union or anti-worker. Instead, he believes that most I/O psychologists are genuinely interested in promoting human welfare through the application of psychology to the world of work and that the goals of the worker and the organization can be integrated for their mutual benefit. He also believes that labor unions are caused by bad management. Thus, unions are not the problem, but a symptom; and, if I/O psychologists can improve the quality of management sufficiently, there will be no need for labor unions, and human energy can be directed to more productive efforts.

Employee selection (both entry-level and promotional) is an emerging problem area which provides a good example of unnecessary labor-management conflict which is destructive to both the organization and the worker.

Seberhagen's presentation discussed eight possible union demands which constitute threats to good practice in employee selection and promotion.

1. Unionization of confidential employees. If personnel selection specialists, test administrators, and other employees directly involved in the development and use of selection procedures are allowed to join unions, there is a serious potential for breaches of test security.

2. Vertical bargaining units. If vertical bargaining units (e.g., all positions in a given department or facility) are permitted, rather than requiring all units to be horizontal (e.g., all positions in a given job class, regardless of department lines), personnel procedures could be adversely affected in a number of ways, involving job analysis, position classification, selection procedures, and transfer of employees within the same job class.

3. Restrictions on recruitment. Seberhagen stated that recruitment is an important, but often overlooked, part of the selection process. If a union negotiates artificial restrictions on recruitment, such as limiting recruitment to certain departments, jobs, or progression lines, not only will the quality of the applicant pool suffer, but also a terrific strain will be put on entry-level selection procedures to assess worker characteristics needed for higher level jobs in addition to those needed for the entry-level job itself.

4. Use of invalid selection procedures. In response to invalid and subjective selection procedures used by management, many unions have negotiated for other selection procedures which they feel are more valid, or at least more objective, to ensure fairness to all candidates. The most common selection procedure advocated by unions is, of course, seniority, but unions have negotiated for other selection procedures, too, such as performance ratings and particular tests. When collective bargaining agreements specify certain types of selection procedures, it is often difficult to establish new selection procedures which are valid because everything is subject to negotiation.

5. Restrictions on test use. Even if valid selection procedures are used, the utility of these procedures may be unduly limited if the union negotiates passing points, ranking schemes, or other forms of test interpretation which negate the value of the selection procedures.

6. Unlimited retesting. Many union contracts require open bidding on all jobs by eligible candidates. This is fine in itself, but there can be a problem if position vacancies in the same job class are announced one-at-a-time. The result is that candidates are either over-exposed to the same test, thereby reducing the validity of test interpretations, or else the employer must go to the great expense of developing a large number of alternate forms of the tests for each job class.

7. Union participation in test administration. Many unions are suspicious of the way management administers and scores its selection procedures, perhaps as the result of past errors or even intentional manipulation of the testing process by management testing staff. In response to these suspicions, some unions have negotiated the right to have a union representative assist in all phases of the administration and scoring of tests and other selection procedures to ensure that everything is done on the up-and-up. Under such circumstances, even if there are no actual breaches of test security by union test administrators, the mere threat of these problems can prevent management from making a major investment in test development.

8. Union review of test questions and answer keys. Even if union members are not test administrators, some unions attempt to gain direct access to test questions and answer keys to ensure that all testing is fair and proper. This was the issue in the case of Detroit Edison v. National Labor Relations Board which was finally decided by the Supreme Court in 1979. Detroit Edison provided validity evidence and other technical data to the union and was willing to show test questions and answer sheets to a qualified psychologist representing the union, but would not give test questions and answer sheets directly to the union. The court ruled that the employer's approach would meet the legitimate demands of the union, while protecting the employer's necessary concern for maintaining test security.

The solution for the eight problem areas just described is not the abolition of unions or tighter restriction of the collective bargaining process. The real solution is proper development, administration, and use of employee selection procedures. The first step in achieving this objective is to educate managers and labor negotiators about good testing practice and the utility of valid selection procedures.

Areas of Concern from the Standpoint of a State Union

Peter Benner, Minnesota State Employees Union

Mr. Benner presented four points of favor on the part of the Minnesota State Employees Union. These included:

1. Seniority is a legitimate factor in the selection process. With other items relatively equal, seniority should be the deciding factor in an appointment.

2. In merit system controlled jobs, unions favor broad eligible lists. They have supported a "Rule of 10" on promotional positions and "Rule of 20" on open-competitive positions. Given the reliability problems and validity problems of written tests and experience and training exams, the assessment procedure cannot give true rankings of candidates.

3. Promotional ratings should be grievable. They are under union contract with the state.

4. While the internal working of selection procedures are normally considered "inherent managerial rights," unions have a legitimate concern in the nuts and bolts. The type of examining procedure, whether jobs are promotional or open to the public, the length of eligible lists, and so forth, can determine whether or not an employee represented by the union has an opportunity for promotion or change of career. The union's concern is with selection procedures which effectively screen-out qualified current employees. Some of these issues should be addressed through the bargaining process, others should be addressed through the legislative process. In either case, the union has a legitimate reason to be involved.

KEYNOTE ADDRESS

The Validity of Content Valid Tests and the Basis for Ranking

Dr. John J. Hunter, Professor, Department of Psychology
Michigan State University

Plantiffs in testing cases have argued that content valid tests should not be used for rank ordering applicants, but instead should be used only to identify those people with minimum competence to do the job. Past that point of minimum competence there would be no further gains in job performance. Dr. Hunter stated that this theory actually postulates a level of maximal competence, and that once this level is reached, there would be no further improvement in job performance. Furthermore, this maximal level of competence is usually set operationally to be so low that 85 percent of all workers would be at that level.

This theory contradicts the major findings of every discipline that has studied work. According to this theory, psychologists using work sample tests to measure job performance should have found that 85 percent of the workers had a perfect score, while the remaining 15 percent trailed down from this point. Instead, psychologists have typically found a normal distribution of performance. The same picture emerges from the reports of policy experts who study production records, management experts who study job evaluation, and administrators who interact with supervisors.

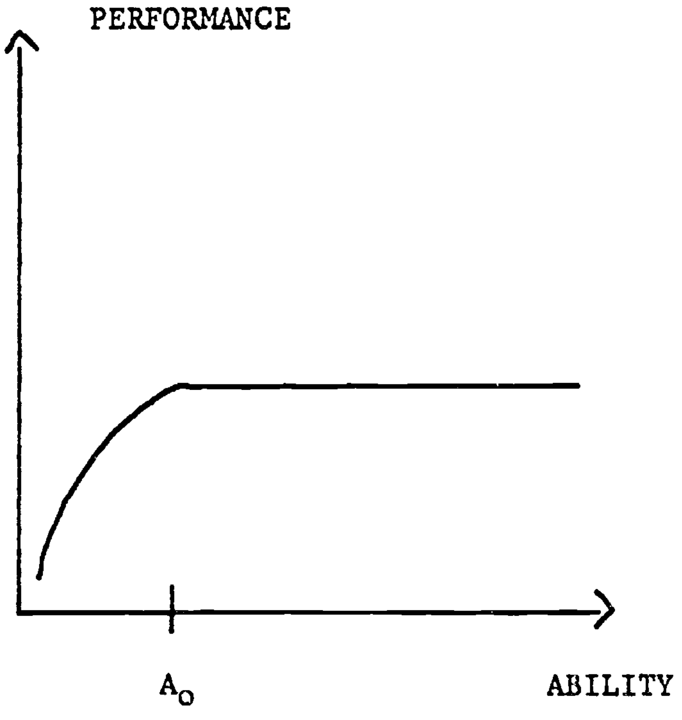

Figure 1a. The nonlinear relation between ability and performance predicted by the theory of maximal competence.

A_o= Level of maximum competence

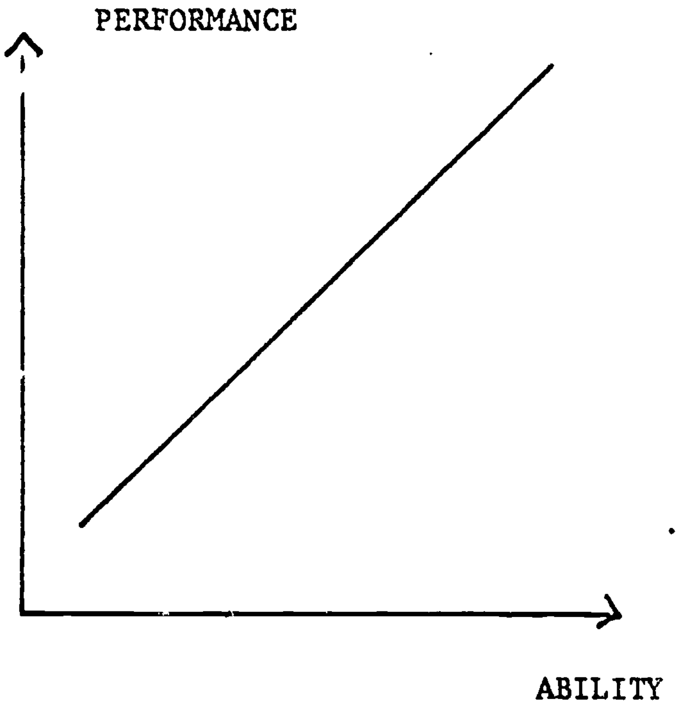

Figure 1b. A linear relation between ability and performance.

Dr. Hunter presented two arguments showing the theory of maximal competence to be false. The first argument related the theory of maximal competence to the judgments made by job experts in the content validation process. Assume that the job knowledge test has 100 items, with a mean score of 50, and that the plantiffs argue that 50 represents the level of maximal competence. Yet a person scoring 50 and a person scoring 100 cannot be equal in job perfcrmance, according to the judgments made during content validation. Job experts defined the tasks that make up the job, identified the critical tasks, and assessed the knowledge required to carry out each task. Thus, each item on the test is linked to some specific type of knowledge required for the performance of some task. Only if all of the job experts were wrong about 50 items on the test could it be that the person scoring 100 does no better on the job than the person scoring 50.

In the second argument against the theory of maximal competence, Dr. Hunter presented cumulative evidence showing job performance to be linearly related to ability test scores. The theory of maximal competence predicts that increasing ability leads to increasing performance only to a certain point. Past that point, everyone is at maximal competence, and the performance curve would be flat. Therefore, the performance curve would be nonlinear.

A statistical test can be run to assess the relationship for nonlinearity in any data set where the ability studied is relevant to job performance. If the relationship is always linear, then evidence of nonlinearity would be found only in chance cases; about five percent of the studies done. Hunter and Schmidt (in press) and Schmidt, Hunter, Muldrow, and McKenzie (1979) found six cumulative studies covering thousands of data sets. The reports were all uniform: evidence of nonlinearity occurred at exactly the chance level. There is no nonlinearity in the employment domain. If a test is a valid predictor of job performance, then the higher the test score the higher the expected performance.

In the recent Guardians decision /630 F.2d 79 (1980)/ the appellate court ruled that a test can be used for rank ordering only if the test is so reliable that a one point differential has a 95 percent confidence value. Dr. Hunter presented a mathematical argument proving that no matter how long the test, the confidence level can never be brought down to one. No test can meet the requirements for ranking suggested by the court in the Guardians case.

No test is perfect, and there will be some misidentification of persons by observed test scores. However, a test can still be useful in selecting those who are most likely to have high job performance. The misidentification resulting from test unreliability has only a small impact on the average job performance of those selected.

Dr. Hunter presented the evidence from the Guardians case showing that the region of major misidentification is the true score region 92-95. This is a very small region in comparison with the entire score range. The benefit of the test is that it rejects those with true scores below 91, and accepts those with true scores above 96. Furthermore, those in the region of misidentification have nearly identical scores on true job knowledge, which means that they would have nearly equal job performance. Finally, shifting from ranking to a low cutoff score would not eliminate the problem of misidentification. It just shifts the region to a different location on the distribution.

Optimum productivity of those selected is obtained by ranking on the basis of the test score, and selecting from the top down. Any departure from ranking will result in lower productivity and, hence, higher costs. The utility equations for the conversion of information about test selection into estimates of the dollar impact on productivity have existed for 30 years. These can be used to compute the dollar losses resulting from the use of any given low cutoff score. The data presented by Dr. Hunter show tremendous dollar losses if low cutoff scores are used instead of ranking.

The use of a low cutoff score has been recommended because it results in a higher selection rate for minority workers. However, it is an economic disaster because so many more low productivity non-minority workers would be selected. Furthermore, the random selection of candidates above a low cutoff score would mean that only a small percentage of the top candidates would be likely to be selected. Many top candidates would not be selected, which is enormously unfair as well as economically catastrophic.

PAPER SESSION

Technological/Methodological Issues in Selection

Chair: Robert H. Marshall, City of Milwaukee
Discussant: Ronald A. Ash, University of Kansas

A Domain Sampling Model For Item and Test Calibration, Analysis, and Equating

Ollie A. Jensen, Educational Testing Service

This presentation dealt with the development of indices, formulas, and procedures that are optimally useful for calibrating, analyzing, and equating items, test segments, and tests used for employee selection or for professional licensing or certification.

Regardless of the specific occupation, there is a finite body of occupational behavior to be tested. This body is further divided by level and kind into separate jobs or into configurations of tasks on which minimum competency must be demonstrated if specified public interests are to be provided specified protections.

Usually, in the set of specifications for a job proficiency test, the general domain of interest (the specified job or the relevant configuration of tasks) is divided into subsets, into test segments, such that each segment measures a specific domain of interest, a separate job function or dimension. Each test segment, each sample from a specified domain, represents for the target population (but not necessarily for the general population) a homogeneous, behaviorally meaningful unit--a particular category of behaviors.

Each item in a category elicits one or another of the behaviors in the domain. The items vary among themselves in such characteristics as variance, means, and covariance with other items in the segment. Any description of the items sampling a domain involves statements about the distribution of item characteristics.

For this purpose, each test item or exercise is described in terms of three statistics (three data points):

p--the proportion of a total group that obtain an acceptable item score (e.g., for items scored right or wrong, the proportion answering an item right);

M_R/n--the unit mean score on a segment consisting of n items for those individuals obtaining an acceptable item score, e.g., those answering the item right;

M_W/n--the unit mean segment score of those obtaining an <u>unacceptable</u> item score, e.g., those answering the given item wrong.

Each test segment is described in terms of three comparable statistics: The mean of the n "M_R/n's" for a segment, the mean of the n "M_W/n's" for a segment, and the mean of the n "p's" for a segment (where mean p equals the unit mean score for the segment, i.e., p = M/n). The total test composed of n_t items) is also described in terms of three comparable statistics: The mean of the n_t "M_R/n's" for the total test, the mean of the n_t "M_W/n's" for the total test, and the mean of the n_t "p's" for the total test.

From these basic statistics the following are derived:

1. An additive (ratio) internal-consistency index that indicates the degree of relationship between item scores and segment scores or between item scores and total test scores relative to the maximum possible relationship between item scores and segment or test scores.

2. Formulas for estimating from the M_R/n and M_W/n values obtained from a particular sample from the target population the probability that a person with a given ability level on the domain will obtain an acceptable item score, and a definition of item difficulty in terms of the probability that a person with a domain score that is at the midpoint of the maximum range of scores on the domain will obtain an acceptable item score.

3. Procedures for determining the existence and extent of differential item difficulty and differential item-segment internal consistency.

4. Formulas for estimating the range of ability on the domain to which the obtained internal-consistency index value applies and the ranges, if any, over which it decreases from the obtained value to zero.

5. The proportion of the obtained internal-consistency index that is due to item self-correlation.

6. A systematic method for setting a cut score in the absence of an external criterion (in the absence of a direct tie between test performance and a specific task performance requirement).

7. Formulas for estimating the skew and kurtosis of a segment or test distribution that are domain-referenced rather than group-referenced.

A detailed application of the statistical procedures (methods) involved was presented.

SYMPOSIUM

Development of Job Related Minimum Qualification Requirements--Methodological and Administrative Concerns

Chair: Marianne Bays, Prudential Insurance Co.
Discussant: David T. Larson, New Mexico State

Update on the Use of the Behavioral Consistency Approach to Unassembled Examining in the Federal Government

Karen Olivia White
Office of Personnel Management

To date, over 20 Behavioral Consistency achievement-based examinations have been developed for use in Federal selection. Several of these have not yet been implemented due to the Federal hiring freeze which has been in effect in many agencies for the past two to three years. The data presented described the experiences Federal examiners have had so far in using those examinations which have been implemented.

The Civil Service Reform Act of 1978 provided for the delegation of competitive examining authority from the Office of Personnel Management (OPM) to other Federal agencies. Several agencies received this authority for specific jobs. Many of the agencies chose the Behavioral Consistency Approach (BCA) for their new examining procedures. Because of the hiring freeze, the only agencies which have had extensive experience using the new exams are the Federal Deposit Insurance Corporation (FDIC) and the Federal Home Loan Bank Board (FHLBB). The exams were developed for Bank Examiner (FDIC) and Savings and Loan Examiner (FHLBB). Both examinations were developed for the entry level.

Both agencies report receiving 400-500 applications each time they have opened their registers for receipt of new applications. The FDIC had hired 117 applicants as of May 1980 and FHLBB had hired 40 applicants as of February 1981. Neither agency has had significant hiring activity since then.

Both agencies report that applicants seem to be screening themselves out, but since both agencies require extensive travel in their jobs, the length of the supplemental form is not likely to be the major cause of applicant self-screening. Both agencies reported that the number of applicants was quite sufficient to fill their needs.

The FDIC noted an unusually low percentage of ineligible applicants after their first open period which was in May 1980. They attribute this to the fact that the examining procedure provides a realistic job preview and applicants seem to pay closer attention to the minimum qualification requirements of the job before they go through the process of completing the application forms. The FDIC also noted that when using the previous selection procedure, the Professional and Administrative Career Examination (PACE), they had an extremely high declination rate.

Both agencies report that the majority of applicants are completing the forms correctly and with sufficient detail. The time required to rate averaged between 15 and 20 minutes per application. Score distributions are essentially normal with a very narrow middle.

Neither agency had received any formal complaints from applicants though a few applicants "grumbled" about the length of the forms. Neither agency had received substantive complaints from selecting officials. The FDIC noted that fewer applicants referred under the new system have a B.A. degree, whereas, virtually all PACE referrals had B.A. degrees.

A comparison was made of selection data by race, ethnicity and sex collected for the PACE in 1978 compared with selection data collected by the FDIC from May through December of 1980 (the FHLBB has made too few selections at that time to determine selection rates fairly).

Data indicated a higher percentage of selection of females, Blacks and Hispanics. The impact ratios for PACE and for FDIC Bank Examiner were also improved for female/male, Black/White and Hispanic/Non-Hispanic.

One factor to be considered in comparing these data is the difference in minimum qualification requirements between the PACE overall and the FDIC Bank Examiner procedure. The FDIC procedure requires a minimum of 24 semester hours or equivalent in business administration, finance, economics, or accounting with at least 12 semester hours in accounting subjects in addition to the qualifications required for PACE eligibility. This minimum qualification requirement serves to lessen the impact of veterans preference against women which might account for the improved selection ratio shown for women.

The most promising aspect of the procedure is the very real possibility that the BCA achievement-based rating offers greater opportunities for women and minorities to be referred and selected. The improved impact ratios seen for women, Blacks, and Hispanics are a result, in part, of the delegation of examining authority; however, the selection procedure itself must take some of the credit. Differences in the quality of candidates overall are not readily apparent; a criterion-related validity study would be needed to detect further distinctions.

Training and Experience Evaluation for the Automotive Mechanic Classification

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

This presentation described the development of a Training and Experience Evaluation plan for the automotive mechanic classification in the Port Authority of New York and New Jersey.

The Port Authority of New York and New Jersey, an organization of over 8,000 employees, is staffed, in part, by numerous trade classifications including plumbers, carpenters, electricians, and automotive mechanics. Entrance into the organization, as well as promotion to higher level classifications, is based on merit principles, keynoted by its personnel testing system.

In the case of automotive mechanic, which is typical of most trade groups, qualified applicants are required to successfully complete a job-related, technical written test. Successful candidates then progress to a technical performance test. Candidates who are successful on both instruments are ranked, based on an average of their scores on both instruments.

Prior to an applicant being scheduled for the written test, he/she is screened for meeting education and experience requirements. In the case of the automotive mechanic applicant, a high school diploma or a GED certificate, and three years of full-time, paid work experience in the technical field are required.

Numerous problems existed with the Port Authority's experience requirement which, while simplistic and straightforward, was also unbending. The following were typically encountered:

1. A candidate with an academic high school diploma and three years experience could be screened into the testing process, while an applicant with a four-year vocational diploma in the automotive field with only two years of experience was screened out of the process.

2. Verification of work experience by the applicant required his/her providing a letter on official letterhead by past employers indicating that the applicant had indeed worked full-time in the automotive field. On occasion, candidates maintained that while they had indeed worked for a given company, the company was no longer in business and, therefore, could not provide written confirmation of employment.

3. While the work may have been in the automotive field, no indication of the breadth or scope of experience was indicated.

4. Candidates may have had experience, but not full-time experience or paid work experience.

5. Letters may have been written by relatives or friends of the candidate who exaggerated the candidate's length of employment.

6. Candidates legitimately may have had three years of full-time paid work experience, working for companies such as Midas, Amoco or Sears Roebuck, where their experience was limited to repetitious work in a narrow field of automotive work such as muffler replacement, transmissions, or battery and shock absorber replacement.

The rationale behind any experience/education requirement or minimum qualification standard is to screen out candidates who are unlikely to be successful in the testing pnases of the employment pıocess. Conversely, candidates who meet minimum qualification standards should, as a whole, perform reasonably well in formal evaluations of their knowledges, skills and abilities. Testing is an expensive and time consuming activity both for an organization and for applicants themselves. The elimination of obviously unqualified applicants is, therefore, a desirable and cost-effective objective.

An analysis of applicant flow into the automotive mechanic classification during the calendar year 1981 indicated that the process as it existed was cost-ineffective. Of some 174 applicants who met education and experience requirements and who were invited to take the initial written test, some 132 or 76 percent appeared. Of these only 34 (or 26 percent) were successful. Of these, 33 appeared for a performance test, however, only three (or 9 percent) were successful in that evaluation.

Since the development of the testing instruments indicated a high job-relatedness, and had correlated well with subsequent performance in the organization, attention was focused on the existing minimum qualifications.

Mr. Forte reported that the result was the development of a Training and Experience (T&E) Evaluation based upon applicant self responses to a questionnaire containing 67 questions. In order to ensure that relevant questions were asked, it was determined that the basic source document for T&E development would be the most recent job analysis report for the classification. Working from the job analysis, a detailed T&E was developed. Knowledges, skills and abilities, which were necessary at time of entry into the classification, were selected for inclusion in the T&E. The questionnaire covered Experience in Preventive Maintenance, Experience in Trouble-shooting and Diagnosing Problems, Experience in Repairing and Replacing, and Experience in Using Tools and Equipment.

Mr. Forte's presentation did not report actual use in the employment process since, at the time of the report, union agreement to delete the existing experience and educational requirements for automotive mechanics and to substitute a training and experience evaluation had not been reached. The author stated it is likely that the union will have to be convinced that a self-completed T&E can be a functional instrument which is capable of assuring an appropriate skill level and is in fact more relevant than traditional minimum qualifications.

SYMPOSIUM

Validation of Selection Procedures With a Small N: An Example of a Successful Approach

Lila Lewey and Richard Olson
Personnel Decisions, Inc.

The technical feasibility requirements of criterion-related validity have led many employers and personnel speciali3ts to avoid it, thereby forfeiting objective selection procedures and the use of standardized instruments. This study presents the results of a strategy for validation that responds to the practical limitation which many employers face in criterion-related validation:

A small n;
Restricted range on test scores and job performance measures; and
Difficulty in obtaining unbiased and reliable measures of job performance.

The strategy was applied for the position of maintenance workers in a midwest corn processing plant. The problems encountered include:

A small n -- only 31 incumbents;
Restriction of range -- practical considerations required the research sample be drawn from job incumbents and so one would expect restriction of range in the selection tests (as with the performance measures); and
The need to rely on supervisory ratings (with all the accompanying potential for bias and unreliability) using job performance as the criterion.

The strategy calls for reduction of random error variance by combining predictors prior to analysis on a unit-weighted basis, rather than using multiple regression procedures. (Regression weights would necessitate cross-validation undoing all of the work to reduce the required sample size.) The unit-weighted predictor composites will require smaller sample sizes than the conventional multiple regression procedure to achieve the same statistical power. In addition, cross validation is not necessary, since there is no statistical capitalization on chance.

The strategy also calls for the other source of contaminating variance -- examining the irrelevant, but systematic, variance resulting from criterion unreliability and range restriction. The three sources of error are interdependent; the more unrestricted the range and the higher the criterion reliability, the smaller is the required sample size.

In this study, open-ended job analysis interviews were combined with an objective Abilities and Work Habits Checklist. The checklist contained both <u>task statements</u> which focused on job-oriented content:

> "Visually inspects parts, components, or instruments for proper fit, alignment, part orientation, or damage," and

<u>work behavior statements</u> which focused on worker-oriented content:

> "Avoids short cuts; has appropriate tools ready and prepares for a job in advance."

Job analysis information (both open-ended interview comments and tabulated results from the Abilities and Work Habits Checklist) was used to develop a list of eight dimensions of job performance, seven specific dimensions and one global measure of overall effectiveness.

The maintenance worker dimensions include:

TECHNICAL ABILITY

1. Skill in machine maintenance and repair
2. Troubleshooting

WORK HABITS AND INTERESTS

1. Thoroughness and accuracy
2. Work energy and initiative
3. Work relations
4. Work organization
5. Training and development of job expertise

These criteria all represent important or critical work behaviors according to job analysis information. Using these criteria, job performance rating booklets were prepared. Performance ratings were then collected and analyzed for quality. Considerable data were presented assessing:

1. The extent of interrater agreement in the job performance ratings; and
2. The raters' relative success in avoiding common rating errors.

The battery of tests selected to predict job success included:

1. Interest inventories;
2. Workstyle measures (personality inventories); and
3. Skill and ability measures.

The complete test battery was administered to 31 incumbent maintenance workers at the plant for whom job performance ratings had been collected.

Three predictor composites were drawn from the initial targeted sample. The composites predicted:

1. Technical ability;
2. Work habits; and
3. Work interests

Each of these was hypothesized to be useful in accounting for some unique criterion variance in the study.

Each of the predictor composites showed validity in predicting overall job success. Combined on a unit-weighted overall predictor basis to form an overall composite, the predictor composite correlated:

.74 with ratings of work habits;
.68 with ratings of technical ability; and
.66 with ratings of overall effectiveness.

SYMPOSIUM

Ethical Considerations in Personnel

Co-chairs: Michele Fraser, Personnel Decisions, Inc., Minneapolis
Robert H. Marshall, City of Milwaukee

Presenters: William Schofield, University of Minnesota
Thelma Hunt, George Washington University
Fay Walther, City of Ft. Worth
Lance Seberhagen, Seberhagen and Associates
David Friedland, Friedland Psychological Associates

This symposium was conducted for the purpose of stimulating IPMAAC members to think about and to generate ideas about what IPMAAC should do with respect to adopting a code of professional ethics.

Opening remarks by the co-moderators briefly reviewed the "rather checkered history" of IPMAAC's attempts to develop a code of ethics and highlighted the heated differences of opinion that had surfaced with respect to the specific form the code should take. The symposium consisted of four major segments:

1. Prepared presentations by Dr. William Schofield and Dr. Thelma Hunt. Dr. Schofield, Professor of Psychiatry, Psychology, and Public Health and Chief of Psychological Services at the University Hospitals at The University of Minnesota, has been on the ethics committee of the Minnesota Psychological Association for the past eight years. Dr. Hunt, Professor Emeritus of Psychology at The George Washington University and Director of the Center for Psychological Service in Washington, D.C., has long been one of the most active and respected members both of IPMA and of IPMAAC.

2. Panel discussion by Fay Walther, Lance Seberhagen and David Friedland. The panel members discussed the kinds of ethical problems they have confronted in their work, and what they would like from IPMAAC with respect to a code of ethics.

3. Small group discussions by the members of the audience, with the aim being that of reaching consensus or identifying major issues to be resolved with respect to the desirable nature of an IPMAAC code of ethics. And, finally:

4. Report-backs by small groups and open discussion. Briefly outlined below are some of the highlights of each of the major segments of the symposium.

Prepared Presentations

Dr. William Schofield

Dr. Schofield began with the broad question: "Why should a professional association of this sort have any concern with ethics?" He pointed out that there would be no need for a code of ethics *if* all members of the profession were circumspect, compassionate, fair, honest, judicious, responsible, trustworthy, and upright -- *if*, more simply, all members observed "the golden rule" or Emmanuel Kant's categorical imperative that "nobody has the right to do that which, if everybody did it, would destroy society." Morality is the over-arching concept which embraces the notion of professional ethics, with professional ethics being specific principles relative to the activities of individuals who have been given, by society, the right and privileges to practice in a special domain. Frequently, professional associations require prospective members to provide character references and, without exception, every statutory licensing bill which licenses a profession requires applicants to give evidence, through references, that they are "of good moral character." His opinion is that professional codes of ethics have evolved out of recognition of basically two facts: (1) Not all persons are consistently moral; and, (2) The practice of any profession provides special opportunities, and special temptations, to behave in immoral ways.

Two major distinctions have been recognized as important by other professional associations:

1. Ethical Principles vs. Ethical Codes of Conduct; and,
2. Professional Ethics (Principles and/or Codes of Conduct) vs. Standards of Practice.

With respect to the first distinction, he made the following points. The essential difference is one of specificity. Principles are expressed in broad terms; Codes of Conduct spell out in detail the particular behaviors that are considered acceptable and unacceptable. Historically, the tendency has been for professional associations (e.g., the American Psychological Association (APA)) to begin with brief statements of broad ethical principles, and then for the broad principles to evolve into more extensive and detailed codes of conduct. Broad, sweeping principles are next to impossible to enforce and they tend to be of limited educational value to members of the profession. Dr. Schofield identified three issues that have been dealt with by every code of professional principles (starting with the Oath of Hippocrates): (a) Competence - the professional is not to exceed the limits of his/her established competence; (b) Exploitation - the professional is not to take advantage of his/her clients; and, (c) Confidentiality - the professional does not make casual disclosure of privileged information.

With respect to the second above-mentioned distinction between professional ethics and standards of practice, Dr. Schofield noted the following. Standards of practice refer to established and accepted procedures within the profession (e.g., the appropriate use of technical methods and tools, the setting of fees, etc.); and they deal with what is considered competent practice, rather than what is considered morally right or ethical practice. Thus where a code of ethics would state that "the professional does not exceed the limits of his/her own competence," the actual definition of competence would occur within standards of practice. Incompetent practice may be as deleterious as unethical practice, and there is an equal need to take care of the matter, but not in the context of ethics. This is a grey area, but the suggestion is that incompetence is not necessarily unethical. A professional society needs to be concerned with and to provide a mechanism to afford oversight with respect to standards of practice, as well as to ethical issues.

In concluding his presentation, Dr. Schofield emphasized the following: In deliberating over what IPMAAC should have with respect to codes of professional ethics and practice, the membership needs to consider an equally important consideration -- how the code(s) will be administrated, operationalized, and imposed.

Development of an appropriate, enforceable code of ethics requires very careful deliberation and very wide sources of input. It is important to give all members of the group the opportunity to have a voice in the development and implementation of the code. It is equally important that very careful thought be given to the mechanism for enforcement so that the code is meaningful and operational, and provides for fair and legal application of sanctions.

Reflecting on APA's orientation with respect to codes of professional ethics and standards, the emphasis is on education rather than punishment. The primary orientation is to have a document which provides members with clear and explicit guidelines by which they can know what is proper and improper in the conduct of their professional work. That is, the document should be used primarily for education. Only in the last instance, where there appears to be incorrigibility, should it be used as the basis for imposition of sanctions (the essence of which is the removal of the individual from membership in the association).

That which "AILs" the unethical professional is: Avarice, Ignorance, or Lust. Ignorance is correctable. The other two may not be.

Contention that it is "too early" for IPMAAC to have its own code of ethics raises the question: Do we need to wait until we have glaring examples of malfeasance? For an association of professionals who have tremendous opportunities to be unethical, it is not too early. The need, primarily, is for an educational vehicle to guide members in the practice of their profession.

Finally, following comments by others, Dr. Schofield emphasized that all codes should be viewed as "living documents," subject to on-going revision and refinement.

Dr. Thelma Hunt

Dr. Hunt briefly reviewed the nature and history of IPMA's activities relating to the development of its ethical code. Major benchmarks were as follows. In 1976, a resolution was proposed to establish a committee to develop the code, with the aim being that of standardizing the principles already guiding the activities of members. Support for this resolution was based on the recognition that most professional associations already had adopted codes of ethics, and so it was probably time for IPMA to do the same, plus, the feeling that having a code of ethics would elevate the status of the personnel profession. In 1977, the IPMA Executive Committee, under the direction of President William Danielson, did establish this committee. In July 1978, this committee provided a recommended code of ethics which was subsequently acted upon and adopted by the membership. The resultant code is very brief and general in nature, comprising eight broad principles. Since the code has been adopted, there have apparently been no troubles that have required calling the code into operation -- no reported violations. (The question was later raised, however, as to whether IPMA had established any mechanism through which violations could be reported.)

Dr. Hunt then focused more broadly on questions that needed to be considered in developing a code of ethics, making the following major points.

What is the function of a code? The function is usually two-fold: (1) Protection of the profession (as minor function), and (2) Protection of the client, which is the public for many IPMAAC members. We need to think about whether the function of protecting "the public," as opposed to individual clients (as is the case for most other professions), might require IPMAAC's code of ethics to differ in some ways from those of most other professions.

With respect to the issue of general versus specific, it is important to separate the three types of codes referred to by Dr. Schofield: (1) Codes of Ethics which give general moral principles; (2) Codes of Professional Ethical Conduct; and, (3) Standards of Professional Practice. The three overlap, but the latter two are detailed and specific while the first is broad and general. One major reason for separating the detailed codes from the general ethical principles is that specifics change over time while general principles tend to remain constant.

It also is important to distinguish between professional codes of ethics and personal codes of ethics. Again, the two overlap. But one implication of this distinction might be that professional codes of ethics do not need to be all-encompassing.

Additionally, legal questions are important to consider. If, on the basis of a code of ethics, an association expels a member, the person's professional license, reputation, earning power, etc. may be endangered; hence, the person may sue the association for damages. Recently, an architect, expelled from the American Association of Architects on the grounds of ethical violations, was awarded damages of $75,000; this illustrates the potential risks of enforcing a code.

Finally, and most broadly, "Do we (IPMAAC) really need our own code when we've got IPMA's and those of all the other organizations to which IPMAAC members belong?"

Panelists' Comments

Fay Walther

Ms. Walther emphasized the following five issues as being salient sources of ethical problems for IPMAAC members:

1. Conflicts between line vs. research responsibilities (i.e., between daily operational activities vs. long range research goals) which cause time and resources to be allocated to "putting out fires" at the possible expense of long-range improvements in personnel assessment activities;

2. Political considerations and pressures that are brought to bear on professionals to serve those political considerations;

3. Privacy of information -- individuals' rights to know what data have been collected and how the data will be used;

4. Merit principle vs. social considerations -- need to balance the two; and,

5. The importance of being over-mindful of the impact the work of assessment professionals has on people and their lives -- that assessment professionals are dealing with people, not simply statistics -- and, therefore, there is a need for a blend of compassion and science.

Lance Seberhagen

Mr. Seberhagen focused on describing the history of the code of ethics in IPMAAC. Major benchmarks included the following. In 1979-80, the IPMAAC Professional Affairs Committee was assigned responsibility for preparing, for the Board, a draft code of ethics. This 1979-80 Preliminary Code of Ethics combines both broad principles and specific behavioral guidelines. The IPMAAC Board rejected this draft as being too specific and asked the 1980-81 Professional Affairs Committee to draft a more general version of the code. The 1980-81 committee subsequently did submit to the board a draft of a brief and general "Code of Professional Principles" which, in 1981-82, was published for full membership review and comment.

Mr. Seberhagen expressed disappointment at the Board's failure to circulate to the full membership the code drafted by the 1979-80 committee. He emphasized that the major difference between the 1979-80 and 1980-81 codes is that the former provides specific, behavioral guidelines, while the latter (like the IPMA code) does not. He judged the former to be substantially more enforceable and to have substantially greater educational value than the latter. Finally, he pointed out that the major lesson to be learned from the Architect's case raised by Dr. Hunt is not that enforcement of codes of ethics is risky, but that codes need to be written to be consistent with local, state and federal law.

Mr. Seberhagen stressed the extent to which the work of assessment professionals has impact on people's lives, and the opportunities that exist for abuse; and, strongly encouraged IPMAAC members to move quickly to adopt ethical guidelines which are sufficiently specific to meet members' educational needs.

David Friedland

Mr. Friedland opened his discussion by further emphasizing the need to ensure that codes of ethics and professional practice are in conformance with local, state, and federal laws; and, by mentioning examples of areas in which professional codes do frequently seem to conflict with current law (e.g., with respect to setting fees, putting prohibitions on advertising, etc.).

He then focused on the need to recognize the differences between IPMA/IPMAAC and most other professional associations (e.g., APA), and to consider carefully the implications of those differences for the development of an effective code of ethics. The differences he cited included: the extent to which the members of the association who belong to clearly defined professions have a clear professional identity; whether members tend to function independently, providing services on an independent basis versus (as in IPMAAC) to work within an organization; members' typical levels of decision-making authority and control over their own professional activities; and the degree of pressure to which members are subject. He noted that until the constituency of IPMAAC can be well defined, it may be premature to go beyond a general code of professional principles. Similarly, it may be inappropriate to state specifically what members should and should not do in particular situations, unless members can be assumed to have relative freedom of action in those situations.

He briefly mentioned a few ethical issues that he considered to be of particular importance. Those were: "red carpet" treatment of those with political connections; the difficulty of maintaining objectivity when called as an expert witness in legal cases, especially when testifying regarding one's own assessment procedures; and, most broadly, the issue of competency -- the extent to which personnel specialists are required to perform many different kinds of work, often without any special training or guidance, and the opportunities this raises for unknowing, unintentional violations of professional standards.

Comments from Audience Groups and Open Discussion

The persons in attendance at the symposium were divided into five small groups (approximately ten persons per group) to discuss the question: "What, if anything, should IPMAAC do with respect to providing members with formal statements of professional ethics and standards?" Subsequently, each small group reported back to the full group. Then, comments were made by individual audience members during the final, open discussion period. Because of shortage of discussion time there was no general consensus developed, but all agreed the opportunity to discuss the many issues was in itself rewarding.

SYMPOSIUM

A Legal and Technical Analysis of Issues Involving the Ranking of Candidates and Setting of Passing Points

William W. Ruch
Psychological Services, Inc.

Keith M. Pyburn, Jr.
McCalla, Thompson, Pyburn & Ridley

The purpose of the symposium was reflected in a carefully prepared eighteen page handout which stated the following:

"The following materials contain leading case law authority and general factual information relevant to the appropriateness and legality of using content valid employment tests to support the selection of employees on a rank order and/or cut-off score basis. The materials are not intended to represent a complete collection of the relevant authorities or factual material, rather these materials are representative of the legal arguments involved in this issue and examples of factual information which can be developed to support the use of content valid tests."

Reference, in the presentation by Mr. Pyburn was made to the early (perhaps precedent-setting) Supreme Court case - Griggs v. Duke Power Co. From this case comes the positive opinion..."Congress has not commanded that the less qualified be preferred over the better qualified simply because of minority origins." Applications of this to the positive supporting of ranking of candidates taking employment and promotional examinations were made.

Two quotations from the Uniform Guidelines on Employee Selection Procedures were presented to program attendees:

"Evidence which may be sufficient to support the use of selection procedure on a pass/fail (screening) basis may be insufficient to support the use of the same procedure on a ranking basis under these guidelines. Thus, if a user decides to use a selection procedure on a ranking basis, and that method of use has a greater adverse impact than use on an appropriate pass/fail basis, the user should have sufficient evidence of validity and utility to support the use on a ranking basis." (Section 5G)

"Where applicants are ranked on the basis of properly validated selection procedures and those applicants scoring below a higher cutoff score than appropriate in light of such expectation have little or no chance of being selected for employment, the higher cut-off score may be appropriate, but the degree of adverse impact should be considered." (Section 5H)

Comments were directed toward the U.S. *Federal Register* publication on *Questions and Answers on the Uniform Guidelines on Employee Selection Procedures*, issued shortly after the issuance of the 1978 Guidelines. Question 62 is: "Under what circumstances may a selection procedure be used for ranking?" The answer is too long to quote, but was supplied to the IPMAAC audience by Pyburn and Ruch. This much is quoted:

"Criterion-related and construct validity strategies are essentially empirical, statistical processes showing a relationship between per formance on the selection procedure and performance on the job. To justify ranking under such validity strategies, therefore, the user need show mathematical support for the proposition that persons who receive higher scores on the procedure are likely to perform better on the job.

Content validity, on the other hand, is primarily a judgmental process concerned with the adequacy of the selection procedure as a sample of the work behaviors. Use of a selection procedure on a ranking basis may be supported by content validity if there is evidence from job analysis or other empirical data that what is measured by the selection procedure is associated with differences in levels of job performance. Section 14C (9); see also Section 5G.

Any conclusion that a content validated procedure is appropriate for ranking must rest on an inference that higher scores on the procedure are related to better job performance. The more closely and completely the selection procedure approximates the important work behaviors, the easier it is to make such an inference. Evidence that better performance on the procedure is related to greater productivity or to performance of behaviors of greater difficulty may also support such an inference." (emphasis added)

The consideration that should be given to linearity of relationship of test (or other selection technique) to the success criterion was discussed. Straight-line type linear relationships lend themselves most readily to rank-order using of selection tests. Linear relationships in which the curve of relationship flattens out at higher test scores may be considered appropriate for lower test score levels, but not for higher ones. A non-linear inverted U-shape curve of relationship between a test (selection technique) and criterion measure is not appropriate for rank-ordering-of candidates use of tests. A quotation presented from the American Psychological Association, Division 14 *Principles* is relevant.

"Selection standards may be set as high or as low as the purposes of the organization require, if they are based on valid predictors... In usual circumstances, the relationship between a predictor and a criterion may be assumed to be linear. Consequently, selecting from the top scores on down is almost always the most beneficial procedure from the standpoint of an organization if there is appropriate amount of variance in the predictor. Selection techniques developed by content-oriented procedures and discriminating adequately within the range of interest can be assumed to have a linear relationship to job behavior. Consequently, ranking on the basis of scores on these procedures is appropriate." (Section B3) Reference was made to specific legal cases to illustrate more specific points and the problems surrounding the cases.

SYMPOSIUM

Productivity, Effectiveness, and Performance Appraisals: Commonality or Confusion?

Chair: Nicholas F. Horney, Salt River Project, Phoenix, Arizona

Presenters: Mark Edwards, Arizona State University
Dennis Sowards, Salt River Project
William D. Hoel, Salt River Project

This symposium addressed the issues of productivity, effectiveness, and performance appraisal by covering these topics: (1) a multi-organizational perspective of productivity and performance appraisal; (2) the process of defining productivity and effectiveness within a public utility; and (3) defining the barriers to productivity improvements within a public utility from a supervisory perspective. Major focus was on the similarities and differences between the concepts of productivity, effectiveness, and performance measure.

Productivity Improvement Through Innovations in Performance Appraisal

Mark Edwards

Edwards stated that productivity improvement through new developments in performance appraisal is based on the assumption that improvements in the measurement process by which organizational rewards are distributed will improve the link between performance and rewards. He presented the procedure used by three large organizations to demonstrate three significant performance appraisal innovations: multiple raters, benchmarks, and rating safeguards. The procedure described demonstrates the use of a participatory process for appraisal development and use. The link between improved appraisal measurement and productivity, and issues needing further research were discussed.

The process for the described productivity improvement is a composite from the results of three organizations: a large public school district, a major international food company, and a large public utility. These three organizations faced similar problems. They were not satisfied with their various performance measurement methods and they could not afford a large time or money expenditure to design and implement a new system. Each believed the appraisal process was the key to linking productive behaviors with rewards. Each organization was willing to innovate to achieve an equitable and practical appraisal system that minimized liability from legal challenges. Furthermore, these organizations wanted participative systems that maximized associate participation in system design and use. However, they did not want a participative process that required too much management time or that required management to abdicate their supervisory responsibilities.

A major concern was the inclusion of organization members in the development of the performance appraisal process that would provide micro (individual) productivity measures. In order to maximize participation, job analysis, criteria development, and appraisal refinement were designed to allow inputs from many organization members. A secondary goal of the process was to require a minimum amount of management time for both the development and use of the appraisal system. Hence, long and involved committee procedures were ruled out.

After job holders, supervisors, and associates were surveyed for the criteria they believed were the most important to productivity and successful job performance, the criteria were combined into 30 factors. Criteria like "writes reports," understood by associates," and "presents summary reports" were combined into a single factor--communication skills.

Each criterion was defined in organizational language so raters would understand each criterion. These definitions also had the objective of ensuring that any criteria used for performance appraisal were understandable, fair, rational, and based on observable behaviors. Knowledge, skills, and abilities associated with each performance criterion were also developed for use in training, development, and other selection decisions.

Raters who participated in the job analysis process included all organization members at the management level. This was done to ensure:

-everyone affected by the appraisal system would have input into its development.

-every potential appraiser would be trained in the scaled comparison rating process.

-everyone would know that the productivity improvement program emphasized individual performance as well as organization performance.

Scaled comparisons were developed for each rater and for each job family the rater rated. The initial desire to weight supervisors heavier than job holders in the rating process was changed to no differential weighting because there was no basis on which to believe supervisors knew the critical job elements better than job holders. Also, rating results from each group were to be identified separately so possible differences between rating groups could be identified.

A single 90-minute meeting was used to introduce the job analysis and criterion development process to raters, to explain the productivity program, and to provide an opportunity to rate the performance criteria. Major objectives of the rater training were to ensure that:

-raters understood the performance criteria definitions.

-they were to rate the job requirements independent of any job incumbent's performance.

-a rater analysis of rating decisiveness and consistency with the group consensus would be prepared on their rating behavior.

The appraisal process includes the use of multiple raters in order to:

-improve the fairness.

-improve the visibility of little known performers.

-improve comparability among many performers.

-allow multiple perspectives of job performance.

-reduce rating biases and dampen the effects of rater errors.

-reduce the "judge" role for supervisors so they can be more effective coaches.

Rater selection required each associate to select three to ten people who have had an opportunity for frequent and significant contact with the associate and who have the ability to objectively rate the job-related performance criteria. Using traditional appraisal procedures, ratee selection of raters would result in a leniency bias that would undermine the performance measurement process. However, leniency and favoritism are dampened in this multiple rater appraisal process because:

-scaled comparisons are used for ratings and they encourage distinctions between performers.

-raters are chosen by multiple "friends" so their ability to rate any one "friend" high is reduced because they have other ratee "friends."

-each rater rates multiple ratees on nine different performance dimensions which encourage objectivity.

-raters understand that individual favoritism is likely to show up in the rater analysis as systematic deviations from the multiple rater consensus.

-each rating decision is a dichotomous choice between two performers or between a performer and a performance standard (benchmark).

Rater selection guidelines provide for a fair and serious rating of each associate's performance on the criteria. Specific ratings by individual raters were not disclosed. This policy helped assure honest reporting and respect for both rater and ratee privacy.

The use of benchmarks in a scaled comparison format improves the effectiveness of the appraisal process. Each benchmark is defined in terms of a performance standard that is typical within the organization. The benchmarks Outstanding, Excellent, Good, and Marginal were used in this example. The benchmark Excellent was defined as "Consistently performs satisfactorily and frequently exceeds satisfactory performance for the criterion." The other benchmarks were defined in a similar fashion.

The results from this micro approach to productivity were measured in terms of acceptance. Two of the three organizations surveyed participants to capture perceptions about the performance measurement system. Both surveys resulted in strong support for both the appraisal process and especially for the use of multiple raters.

Productivity Measurement in a Public Utility

Dennis Sowards

Sowards' paper was designed to aid utility industry managers in identifying and adapting productivity measurement tools for use in improving productivity. Specifically, his paper details a year-long effort by the Salt River Project (SRP), a major water and power utility in Arizona, to tailor a performance measurement and improvement program for its own use. The paper explains the events leading to the establishment of the program and describes the various components of the program. These include several different levels of productivity measurements, along with a discussion of the objectives of each. The remainder of the paper covers the steps necessary to complete the implementation of the entire program at SRP. Sowards' presentation of criteria for establishing productivity measures should be of particular interest to many IPMAAC members.

In developing a measurement system, it is necessary to first develop the criteria upon which the measurement system will be judged. Two main types of criteria were set for SRP's productivity measurement system: criteria relating to the theoretical measures themselves and criteria for the mechanics of the system.

The criteria regarding the theoretical measures deal with the ability of the measures to tell the correct story. The measures must be:

1. Easily understood by upper management, as opposed to a few technical experts.

2. Show changes that actually occur. (Any change in the model results must be traceable to real changes in the system not due to the model itself. The change must be identifiable.)

3. Allow for forecasting future changes or opportunities. (The ability to isolate potential changes in the components and view the inputs make it possible for management to ask "What if?" questions using the model.)

4. Show productivity changes in the ratio of output/input values.

These criteria are not listed in any particular order of importance.

Criteria on the mechanics of the measurement system deal with the practical use of the measures. They are (in no order of importance):

1. The measures must be timely. The data and calculations must be such that reports can be issued quarterl: match the quarterly financial statements.

2. The data must be accurate, matching all other published data.

3. The measures must be consistent and comparable, treating data the same from year to year.

4. The measures must not cost more to measure than they are worth. This includes the cost of maintaining the reporting system itself.

The model to measure productivity was developed by the American Productivity Center in Houston, Texas. The basic idea behind the model is to measure the total of corporate outputs divided by corporate inputs. In other words, it involves a measure of production divided by resources expended. In order to measure the progress of the company from year to year, a base year is used, and then performance in successive years is measured against performance in the base year.

Three separate measures are used in the model to detect changes in corporate performance. The first, called productivity, is a measure of how efficiently the company can translate units of resource inputs into units of production outputs. The second is called price recovery, and measures the rate of change in the price of outputs divided by the rate of change in the price of inputs to the company. The final measure involves profitability which is calculated by multiplying the productivity index by the price recovery index. The result is a measure of overall performance.

Development of the model required several steps. First, the corporate outputs and inputs had to be defined. Once the output and input categories were defined, data on the input usage and output production had to be gathered. While the data were collected, a computer program was being written to do the mathematical calculations necessary to determine the productivity, price recovery, and profitability indices for each year. When the program was finished and data collection completed, preliminary runs of the model were made and the process of interpreting the results began.

The total factor productivity measurement system development project is nearing completion. When the capital issue is resolved, the model will be ready for presentation to executive management.

Productivity, Performance Appraisal and People

William D. Hoel

Hoel's contribution to the symposium emphasized the personal interaction element. He stated that we must look at the human element in the organization when putting in a productivity improvement program. Furthermore, a system that ties pay to performance must be developed, and it must be perceived by the employees. One way to do this is to strengthen the communication process between the employees and their supervisors. Just as importantly as the above, policies and procedures in the organization that run counter to and decrease the motivation to produce need to be changed at the same time. Otherwise, credibility of the program is lost. Employees will say "management really doesn't care and doesn't want this program" or "they ask me to do it on the one hand and don't give me the resources to do it on the other." Remove the perceived barriers to increasing productivity such as lack of skills, knowledge, and abilities. Tie the expected productivity to your appraisal objectives and reward system and you will be well on your way to a more effective, efficient organization.

INVITED LUNCHEON SPEAKER

Public and Private Sector Assessment: Is There a Difference?

Dr. Virginia R. Boehm, Assessment and Development Associates

Differences between public and private sector assessment exist primarily at the micro level and relate to specific techniques and statistical procedures. On broader issues there are fewer differences and more commonalities. Dr. Boehm addressed three broad challenges that are faced by assessment professionals in both the public and private sectors.

The first of these challenges is the need to comply with two competing priorities. Assessment professionals are working to improve productivity and, at the same time, to comply with equal opportunity and affirmative action requirements. When using a valid test that has adverse impact, assessment professionals use various ingenious methods to balance these competing priorities. They add other components to the selection process, set minimal passpoints, use broad score bands, or use separate rankings. These methods are all scientifically without merit and legally debatable.

Dr. Boehm discussed the Teal case, in which it was argued that the fact that the State of Connecticut had made up for the adverse impact of the test in the later stages of the selection process was an ameliorative action. This action was said to be an admission that a wrong had been done, that the test was not fair to minorities, and that the plantiffs had been discriminated against. Dr. Boehm pointed out that it was unfortunate that the State did not defend their actions as a well intentioned attempt to resolve competing priorities. This would have encouraged the court to address this issue in its decision on this case. Dr. Boehm stated that the problems created by these different priorities cannot be solved through pseudo-statistical juggling. It is a political problem, and it must be solved through political means: policy, legislation, and court decisions. Whatever the outcome of the Teal case, hopefully, it will generate more cases which will address this issue more directly.

The second challenge faced by assessment professionals is attempting to maintain professionalism and productivity in an environment of declining resources. Budget cuts, freezes, and RIF's have both direct and indirect impact. Declining budgets and generally bad economic conditions mean that both government and business have greatly decreased the number of people who are being hired and promoted. It is difficult to justify allocating scarce resources to selection when fewer people are being selected. However, these same conditions mean that there will be a higher number of applicants for the few positions that are open, which means there is actually more, rather than less, demand for assessment and selection. In addition to this, declining resources mean that mistakes in selection are less organizationally tolerable. The challenge is to do more with less, and to do it better.

The third and most serious challenge faced ɔy assessment professionals is the compelling need to develop radically new assessment tools. Dr. Boehm expressed doubt that anything further can be done to increase the proportion of individual productivity variance accounted for by paper-and-pencil tests, and she cited three types of evidence to support this statement. Her review of criterion-related validity studies of the 60's and 70's, published in Personnel Psychology, showed that the average validity coefficient failed to change in spite of changes in the types of measures used, the study design, the sample design, the sample size, and the type of criteria. Not only did it fail to increase during those twenty years, it was virtually identical to the average validity coefficient obtained by Ghiselli in a review of studies done in the 40's, 50's, and early 60's. The same conclusions can be drawn from the validity generalization work of Hunter, Schmidt, and others. The fact that validity generalization works so well indicates that there has been little or no improvement in the tools used. A third line of evidence comes from looking at what sells. Many of the most widely used tests are quite old. (These include the Bennett Test of Mechanical Comprehension, the Purdue Pegboard, the SRA Adaptability Test, the DAT and GATB aptitude batteries, and the Stanford-Binet and Wechsler intelligence scales.) The fact that so many of these old tests are still in use is an indication that efforts to develop superior new instruments have not produced significant improvements in the eyes of test users.

Dr. Boehm cited one solution to this problem as further refining two of the newest tools in assessment, biodata and assessment centers. However, she added that this will not be a sufficient response to the challenge. What is needed is a re-examination of the approach to both content and criterion validation starting with job analysis. Current methods of job analysis focus too narrowly on the tasks performed by an individual worker in isolation, and do not consider the impact of the work environment, the role of other workers, motivational factors, and incentives offered in the work place. Yet productivity depends as much or more on these types of factors as it does on the ability of the individual worker to perform a set of tasks.

Dr. Boehm stressed that she was not advocating a return to the use of either personality tests or vocational interest tests in personnel selection. She suggested that job analysis be broadened to include work context factors that can be measured. The "work analysis" would look at not only job tasks, but also such things as the degree of autonomy provided the worker, the length of the work cycle, how closely the job is tied to other jobs, the amount of variety in the work, etc. Then, assessment tools should be developed to measure individual preferences for these contextual factors. These assessment tools would be subjected to the same standards of reliability and validity as aptitude and proficiency measures, and they would be demonstrably work-related and linked to productivity. These tools, in conjunction with aptitude or proficiency tests, should increase the proportion of variance accounted for by conventional tests alone.

Boehm suggested that assessment professionals in the public sector are in a position to take the lead in developing these new assessment tools. They rest on the analysis of work and its context, and the greater standardization of environments should make reliable measurement easier in the public sector. She concluded by pointing out that these challenges faced by professionals in the assessment field will only be met through unified effort, because they require all the expertise of those in both the public and private sectors.

PAPER SESSION

Selection Procedures in Police and Fire Settings

Chair: Karen Coffee, Cooperative Personnel Services
Discussant: Lance Seberhagen, Seberhagen and Associates

Development of the Written Test for the Classification Apprentice Fire and Rescue Officer

Donald A. Emmerich, City of Dallas

Donald A. Emmerich's paper deals with the selection of fire apprentice and rescue officers in the city of Dallas, Texas. It is concerned with the development of a written examination meeting certain problems arising in the use of an earlier carefully developed and validated test. As Emmerich stated in his paper, since the development of the original written test for the classification, Fire Apprentice and Rescue Officer, several conditions have arisen which have dictated a re-evaluation of the selection procedures devised in the original study. The primary condition was the implementation by the Fire Department of a policy which required all Fire and Rescue Officers to attend and successfully complete Emergency Medical Technician (EMT) and Paramedic training programs.

The practical concern on the part of the Fire Department, as presented by Emmerich, related to the large proportion of selected Fire and Rescue Officers dropping out or failing, presumably from the EMT training requirement. This was very expensive in terms of lost human resources and training time, and prompted Fire Department officials to ask whether a more efficient means of selecting officers might be developed. When these circumstances were brought to the attention of the Civil Service Test Analysis Division, it was apparent that this major change in job duties had seriously affected the validity and utility of the selection procedure in use.

Emmerich's paper presented a detailed examination of prediction variables and criterion variables (available on previous selected groups) as such a study might relate to an improved selection procedure (selection test(s)). The prediction variables included: (1) Research Test Battery (written tests with a number of subtests considered separately in aspects of the study); (2) A Physical Ability Test (detailed components of which were not presented in the paper). The criterion variables studied included: (1) Emergency Medical Technician (EMT) Performance; (2) Firefighting Performance; and (3) Training Academy Performance. Significant correlations were reported between the Research Test Battery (written test) and all three criterion measures with higher relationships being indicated for predicting EMT Performance and Training Academy Performance than for Firefighting Performance.

Emmerich did detailed studies of the interrelationships of the prediction and criterion variables, and presented in his paper multiple regression equations for predicting the three criterion measures from the detailed components of the two basic prediction measures. The Physical Ability Test (PAT) showed less predictive value in the relationships studied than did the written test (Research Test Battery). However, significant relationships were found between some of the PAT sub-test variables and Firefighting Performance. Two variables from the PAT dummy carry exercise sections which were included in the final recommended list of selection tests. The recommendations for tests (test parts or components to be used for prediction of each of the three job success criterion) were based upon the predictive regression studies.

Finally, of the tests studied, Emmerich recommended that the following be included in a selection testing procedure: The Otis Lennon Mental Ability Test (included in the Research Test Battery), seven scales of the California Personality Inventory, the following parts from the written Research Test Battery (Dots, Clerical, Arithmetic. Spatial, Maze, and Math), and the two parts of the Physical Ability Test having to do with the dummy carry exercises.

Emmerich's paper had some other practical comments on presenting test results of applicants, recommending "group" ranking rather than individualized ranking.

Development of a Fire Simulator to Select Chief Officers in the Fire Service

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

Patrick T. Maher's presentation focused on the development of a fire simulator utilized in the promotional process for chief (high level) officers in the fire service in Minneapolis, Minnesota.

The fire simulator presentation consisted of a description of a high rise fire fighting situation in an office building. A most interesting aspect of this report was the description of the nature of the "simulator." It consisted of a very innovative combination of verbal, visual (video), and auditory presentations of a fire situation, as might be encountered by a "firechief officer." Events in the fire situation were presented in sequence. During the total presentation the candidate was required to indicate in detail, including verbalizing thinking, all the decisional processes involved in carrying out the management of the available fire fighting resources and personnel.

Maher's choice of the term "simulator" for this promotional device may introduce a new use of the term for many personnel directors. But, let's keep up with the times and add to our definitions.

The "meat" of Maher's paper is two-fold: (1) the setting forth of the content validation of his promotional simulator; and, (2) the development of a scoring or rating process that reduces subjectivity to a minimum.

The more practical criteria of the success of his promotional procedure can be indicated by his own summary statement. Experience has shown that the process is accepted by the department and the candidates. The candidates feel that the process is impartial, job-related, and a test of ability, while department executives are satisfied that qualified candidates are selected.

Critical Incident Oral Examining: A Technique for Improving Fire and Police Promotionals

Janet McGuire, Arlington County, Virginia

McGuire reported the development of oral examinations for Fire and Police promotions in Arlington County, Virginia, utilizing a Critical Incident technique of development. She described a successful attempt to improve promotional systems in both a police and fire promotional context including the use of a work simulation approach and the development of oral examinations based on critical incidents. Use of managers from within the department to develop critical incidents, objective scoring checklists and a set of "checks and balances" resulted in strong internal "ownership" of the process. The oral examining procedure was perceived by applicants as having high face validity; grounding in job analysis data gave evidence of content validity from the perspective of the Personnel Department. The response format was chosen to maximize mirroring of the target job. Candidates were asked to list all the things they would consider in making a decision on resolving problems which they might have to face when promoted. This obviated complaints that divisional or platoon differences in acceptable tactics or work procedures would unfairly penalize individuals in any objective testing procedure.

Fire and Police Department reaction to the procedure was strongly positive. The technique had good interrater reliability and tended to correlate more highly with management ratings of performance than did the written examination which had also been used in the promotional process.

Development and Validation of an Interest Inventory for Police Selection

Bruce Davey, State of Connecticut

Bruce Davey's contribution to this paper session emphasized the importance of attention to non-cognitive testing (evaluation) in the selection of police officers.

To quote from Davey's paper: "Most selection tests for entry-level police officers are cognitive in nature -- basic ability tests primarily measuring the areas of verbal skills and logical reasoning ability. Affective traits, such as personality, motivation, and interest, usually are not measured until later phases of the selection process. Yet job analysis seems to consistently show that these affective areas are extremely important for success in police work, and sound testing practices tell us that the most important traits should be measured at the beginning of the selection process, not at the end." He pointed out that his presentation was not to undermine the importance of verbal and reasoning ability, written communication, and the ability to learn police procedures and the law, factors typically involved in selection tests for police officers.

Davey's presentation, as implied in the title of his paper, has to do primarily with an interest test (inventory) for police officer selection. The inventory aims at evaluating attitudinal, motivational, and "take-charge" qualities among the critical characteristics that appear in police job analyses. Efforts to evaluate such characteristics by the typical "personality inventory" generally have not been used for police selection.

Davey's approach to developing a selection instrument is unique in using a "Q-sort" technique. In fact, he has called this test the Vocational Interest Q-Sort (VIQ). The VIQ consisted of a deck of 27 cards which the test-taker was instructed to rank-order. Each card described an activity or an interest in some detail, and there was a wide variety of activities represented across the 27 cards -- none of which were specific to police work.

Validation of the Interest Inventory involved studies conducted on 246 Connecticut State Troopers. In addition to completing the Vocational Interest Q-Sort, the troopers filled out a self-rating form in which they rated themselves on traits related to motivation. Motivation was the affective complex aimed to be tapped with the Q-Sort. The first phase of this project was to find out what sorts of basic interests and preferences were characteristic of Police Troopers who rated themselves as highly motivated, as being interested in their day-to-work, and as being satisfied with their career choice.

The next step in this project was to convert data from the Q-Sort format into objective items. To accomplish that, interests were paired off with one another. A Q-Sort item that was positively correlated with Trooper self-ratings of motivation was paired off with a Q-Sort item negatively related to motivation. In deciding which items would be paired, Davey observed the time-honored psychometric principle that the item with the most potential for validity is the one at the 50 percent difficulty level -- so in pairing, items were put together which were substantially equal in popularity. Using the two principles of pairing positively predictive items with negatively predictive ones and equating their difficulties as closely as possible, an 18-item objectively scored interest subtest was developed.

Although the interest test was developed with a validity criterion of the Trooper's self-ratings of interest and motivation, Davey presented data indicating positive support for the test in raising correlations of selection tests with job performance ratings when the Interest Test is added to the selection battery.

Davey emphasized that his interest inventory is not the answer for valid and impact-free police tests. The inventory is only one of many tests and taps only one area related to police effectiveness. However, it has promise for increasing validity, while reducing the adverse impact of police selection batteries.

SYMPOSIUM

The Use of Written Simulations in Personnel Selection

Harold P. Brull, Personnel Decisions, Inc.

Work sample tests include such things as assessment centers, role playing, and group discussions. All of these types of tests have limitations in terms of time, expense, and logistics, and are impractical for testing large numbers of people. Written simulators are a type of compromise; they put people into a simulated environment, but on paper.

The written simulation technique was originally developed at the University of Illinois Medical School by Dr. Solomon and Christine McGuire in an attempt to measure competency in students. The developers thought that competency included not only job knowledge but also judgment and application of knowledge.

A job judgment test consists of multiple choice items which require the examinee to apply knowledge in a particular work situation and choose the right alternative. Written simulators take this idea one step further and weave the separate situations together using a multiple branching format. How the examinee responds on a particular item or section determines the next set of problems to be solved. For example, after being presented with the patient's complaint, the examinee faces several alternate courses of action, such as taking an oral history exam or conducting a physical exam. If the examinee chooses to take an oral history exam, the next problem is to choose what questions are to be asked.

The written simulator test can be constructed to include feedback and additional information which might influence the situation or affect the examinee's judgment. For example, patient responses can be included in the test booklet in invisible ink, and uncovered when the examinee chooses to ask a specific question. Other information can be included also. For instance, if the examinee decides to order an X-ray, it can be found in an envelope contained in the test packet.

This type of test measures job or task knowledge along with application of that knowledge. Other traits that are measured include: skill in assessing priorities, skill in eliciting data, skill in using a variety of resources, efficiency in solving problems, decision-making skills, and skill in manipulating the situation to resolve the problem.

The traditional tools cannot be used to analyze item responses on this type of test. Items on written simulation tests are not independent. Each examinee may answer a different set of items because of the multiple branching format. This makes internal consistency very difficult to measure. Brull reported that the information available in the literature on internal consistency is inconclusive and contradictory. The research he found does indicate that test-retest reliability is very high.

Written simulation tests must be demonstrated to be content valid, and to sample the domain covered by the job. In constructing the test it is important to create enough situations that encompass a broad range of the demands of the position. This will make it defensible as a content-valid test.

Sooring is difficult due to the multiple branching format. The literature reports three types of scores. The proficiency score represents the degree to which the examinee's total approach to solving the problem approximates the correct approach as defined by job experts. The efficiency score is determined by a particular section of the test in which the examinees can gather as much information as they deem necessary. (For example, the examinee might have to chooce which laboratory tests to administer to a particular patient from a wide-ranging list. Some of these tests would be necessary and helpful, some would be neutral, and others might actually be risky or harmful to the patient.) Job experts would have pre-determined what pieces of information are the most helpful and necessary. The efficiency score is determined by the number of positive choices as a percent of the total number of choices. The third type of score is the total competence score.

Brull described in detail the development of a written simulation test for positions within the California State Prison System (a project of the California State Personnel Board).

SYMPOSIUM

Professional Accountability to Applicants

Chair: Glenn G. McClung, City and County of Denver

Presenters: Doris M. Maye, State of Georgia
Steve J. Mussio, City of Minneapolis
Janet L. McGuire, Arlington County Department of Personnel
Sally A. McAttee, City of Milwaukee

Glenn McClung explained that questionnaires had been mailed to 63 IPMAAC members who represented different public agencies in order to determine existing rules and policy governing issues of accountability in the personnel field. From the 24 questionnaires that were returned, it was determined that most jurisdictions do have some legislation in these areas, but there is considerable variation in what is considered accessible. Employment applications are publicly accessible in almost as many places as they are held in confidence. The same is true of personnel folders, and to a lesser extent, test results. A large majority of employment application forms still request information on age and criminal convictions, and about half request information on citizenship, sex and race. A prevailing practice is to make test results available to applicants; and, in about half of the jurisdictions, test results also are available to the hiring authority.

Most jurisdictions offer some type of post-test review; slightly more than half allow review of the test booklet (in at least some cases), while the rest allow only the score and answer sheets to be reviewed. Almost all jurisdictions have some method for handling protests and grievances, but only half have formal internal appeals or hearings. Most require a written protest which is subjected to staff analysis. The most common remedies are deletion of items or rescoring of tests.

A number of jurisdictions have regulations or laws making personnel files publicly accessible; however, certain information is considered public while other information is not. Working papers, internal memos, and procedural manuals are generally considered non-controversial and are accessible to the public.

Doris Maye began her presentation by summarizing an article from the American Psychologist that indicated that while user institutions historically had control over the disclosure of test information, test takers are now gaining influence. The consumerism movement is entering the debate on testing issues. Ms. Maye felt that in order to stem the tide of hasty legislation, it is important to give the applicant as much information as possible without endangering test security.

Prior to taking the test, the applicant should be informed as to the type of test question, test format, time limit, general content areas, cutoff scores, and method of determination of final scores. Applicants should have an opportunity, during or following the examination, to express problems and complaints. In the State of Georgia, an inquiry form is available during the examination. It is reviewed, and if a legitimate complaint is found, the item is not counted against the applicant. Feedback is given promptly, and applicants very seldom pursue the matter further. It is also important to give the applicants timely notification of results, and, if possible, to provide a breakdown of their performance in each area or section.

Steve Mussio discussed the existing state law in Minnesota concerning the issue of invasion of privacy. He noted that this legislation, which specifies what data elements may be released to the public, changes almost every year.

Sally McAttee discussed open vs. closed test review procedures. The rationale for test security is that without it, test validation is destroyed. Yet, people have a right to know as much as possible about how decisions are made about them (this objective is part of American Psychological Association Standards). Not allowing applicants to review their tests promotes alienation and mistrust because they have no other way to verify the correct answers and see that their papers were scored correctly.

There is a fear of accountability because it may show up problems within the field of testing. However, openness alleviates mistrust and wards off potential challenges. Accountability also helps to keep professionals on their toes, to make sure that tests are content valid.

Ms. McAttee reviewed the Detroit Edison vs. National Labor Relations Board case, which is often cited as support for test security. In this case, the Supreme Court ruled that the Board had abused its discretion when it ordered Detroit Edison to turn over its test battery and answer sheets to the union. The Court also ruled that the company's unwillingness to disclose scores and names associated with them, without the consent of the employees, was not lack of bargaining in good faith. This case is not generalizable and may not provide support in other cases concerning test security for the following reasons:

1. The company was not making a case for content validity, and so the union had no need to see the test booklet to check the content of the test.

2. This was an aptitude test in which validity is best judged by a psychologist. In more commonly used knowledge or ability tests, the applicant often has some knowledge or experience with which to judge the validity of the test.

3. The company had offered to have the test reviewed by a psychologist appointed by the union (as a gesture of good faith).

4. Privacy was an issue in this case. The union wanted disclosure of all names and scores and the company wanted permission from the examinees to release such information.

5. Detroit Edison had already reviewed the test with each applicant.

Janet McGuire discussed security versus accessibility of test materials. She outlined four distinct situations: promoting outside applicants, promoting inside applicants, using tests only once, using tests repeatedly. Laws and regulations that govern test accessibility do not make any distinctions between these various situations, yet the differences should be taken into account.

Ms. McGuire made the following suggestions for giving applicants the information they need, while maintaining test security:

1 Give realistic job expectations.

2. Give all applicants clear and accurate information concerning all procedures.

3. Give applicants developmental information rather than giving them all the details (they want to know how they can improve their performance).

4. If it is necessary to change a procedure, tell all applicants about the change and explain the reasons for it.

5. After the procedure is over, ask the applicants what they thought of it and how they would change or improve it.

PAPER SESSION

Performance Appraisal and Evaluation

Chair and Discussant: Nicholas P. Lovrich, Jr.
Washington State University

Performance Appraisal: New Directions for the 1980's

Nicholas F. Horney
Human Resources Department
Salt River Project

Little research attention in the performance appraisal literature has been given to the effects of the perceived purpose of performance appraisal, perceived fairness and accuracy of ratings, or trust in the appraisal process. Although it has been hypothesized or shown how each of these may affect performance appraisal ratings, the extent to which variables such as these interactively affect appraisals has not been researched.

In October and November 1981, approximately 100 managers and supervisors completed a composite of several instruments as part of a performance appraisal system training program. They responded to a modified version of the Trust in the Appraisal Process Survey (TAPS; Bernardin, 1978). Employees were asked to indicate on the TAPS their agreement/disagreement, on a five-point scale of intensity, to 13 statements which describe the rating behavior of the "typical" supervisor in their department. Statements on the TAPS describe rater behaviors that could result in inaccurate ratings. As pointed out by Bernardin, Orban and Carlyle (1981, p. 312), the "typical" rater might be seen to "purposely inflate ratings," or "will distort ratings to get a better deal for his/her subordinates." High agreement with the TAPS items implies that the rater feels other raters are inaccurately rating their subordinates (e.g., inflating the ratings) and thus making a rating error.

The sample was to limit only 42 of the original 100 completing the questionnaire. Since the new performance appraisal form, developed for the system, was to be used only for evaluating supervisors (N=600), the emphasis was with the actual ıatings and perceived rating behavior of supervisors of other supervisors.

The results indicate that raters' trust in the appraisal process can influence ratings of their own subordinates. The TAPS measure accounted for a substantial portion of nonperformance related variance in ratings. One potential explanation of the results is that if raters believe others will inflate their subordinates' ratings, for whatever reasons (e.g., merit money allocations, job evaluation enhancement, fear of confronting employees with low performance ratings, etc.), then they will inflate their own ratings as well. Therefore, the development of a new performance appraisal form may be necessary but not sufficient for the development of a viable performance appraisal system. Other personnel/human resources' systems should be considered when developing a performance appraisal system.

The failure to find a significant difference between high and low "trust" groups could be partially a function of the small sample size (N = 42). In addition, the restriction of range in performance ratings could also be partially responsible. Further investigations are planned which will include a larger sample size.

Note: This series of papers contained an excellent summary of pertinent literature on performance evaluation, and included a good bibliography. You may wish to write to the authors if this information is needed.

The Development and Validation of Supervisory Appraisals for High Grade Personnel

Steven D. Norton, Ph.D.
Department of Defense Centralized Referral Activity

Lieutenant Colonel Edward J. Dunne, Ph.D.
Air Force Institute of Technology

Two Types of Supervisory Roles: In developing an appraisal system and planning validation studies, it is important to clearly describe and define the target job. The relationships of a supervisor with other organizational members can be viewed as being one of two basic types.

The Working Supervisor: The first basic type of supervisor is the "working supervisor." He or she is expected to know more about the details of the job performed by subordinates than do the subordinates themselves. The primary capabilities required in addition to technical job knowledge are human relations skills and knowledge of whatever administrative procedures are required of first-level supervisors. In order to know more about the jobs performed by subordinates than do the subordinates themselves, the supervisor must supervise jobs which are low in differentiation (e.g., a typical production line), must have a small span of control (e.g., a supervisory laboratory technician in a hospital), or must have an educational background which subsumes the knowledge required by all of the subordinates (e.g., a doctoral-level research scientist supervising a team of technicians). Note that in each of the situations above, the subordinates act as the "hands" of the supervisor.

Mintzberg (1973) calls the working supervisor a "real-time" manager who operates primarily in the present, devoting effort to ensuring that the day-to-day work of the work group continues without interruption and is prepared to substitute for any subordinate.

For this type of supervisory position, selection and, thus, selection procedures are based primarily on specific job knowledge. The production supervisor who has no managerial responsibilities can be selected from the most experienced and dependable production workers. The supervisory laboratory technician's prime qualifications are a higher level of education, experience, or accreditation than subordinates. The lead research scientist is chosen because of education and technical achievements. For selection procedures, the content/construct distinction is not likely to be troublesome.

The Supervisor-Manager: The second type of supervisor may be called the "supervisor-manager." The supervisor-manager is not expected to know more about the job tasks performed by subordinates than do the subordinates themselves, but is expected to be a "manager" with all this term implies. The "insiders" and "team managers" described by Mintzberg (1973) are supervisors-managers. A supervisor-manager counterpart of the working supervisor laboratory technician described earlier would be a head nurse who managed a department and supervised many specialized nurses. A research scientist who was also a supervisor-manager would supervise other scientists (or senior technicians) who knew more about their fields than he or she did. He or she would, however, coordinate their efforts, develop plans for future research efforts, and interact with higher management in decisions about the work organization.

Selection and, thus, selection procedures cannot be based primarily on technical competence because, for example, leadership and administrative skills may be very important in these positions. Thus, selection should also be based on other less tangible skills and abilities which are often thought of as constructs. The subject of this paper is the development and validation of appraisals to be used in the selection of this type of supervisor -- the supervisor-manager.

It is possible to measure the "constructs" important in selecting a supervisor-manager, because the assessment center is designed to measure these "constructs" and is a powerful predictor of success in such positions. Its strength derives from two characteristics. First, it translates attributes which might be viewed as constructs into observable behavior. For example, "leadership" becomes the ability to convince a group to follow a course of action. "Decisiveness" becomes the ability to make a decision with somewhat inadequate evidence, when it is clear that it would be harmful to avoid making a decision. Second, the assessment center is independent of the candidate's current job duties and the ability and motivation of the candidate's supervisor to provide accurate appraisals (Norton et al., 1980).

The Behavioral Consistency Method: However, assessment center appraisal is quite expensive and frequently not practical. We believe that an adaptation of the behavioral consistency method can increase the validity of appraisals over that resulting from traditional methods.

"The behavioral consistency method is a nontest selection procedure based on the past achievements of applicants rather than on credentials" (Schmidt et al., 1979, p. i.). It is grounded in the well-accepted principle that the best predictor of future behavior of a given kind is a measure of similar past behavior. This principle was used to develop a procedure for filling mid-level federal jobs with candidates not currently employed by the federal government. Although it is described as "content-valid," some of the factors measured (e.g., analytical and quantitative reasoning abilities; interpersonal and organizational skills; and motivation, initiative, and the ability to organize work have been considered constructs in some court cases.

The factors resulting from the behavioral consistency method certainly should meet the Guardians test of being "the most observable abilities of significance to the particular job in question." It follows the approach to validation discussed by Guion (1980) and Cronbach (1980), although they were focusing on paper-and-pencil tests. How successfully this approach can be adapted to supervisory appraisals remains to be seen.

The proposed adaptation of the Behavioral Consistency Method to the development of appraisals for use in high level competitive promotions was outlined in detail.

Executive Evaluation: Assessing the Probability for Success in the Job

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

The Department of Agriculture implemented a performance appraisal system for its executives which was based, at least in theory, on objective, job-related performance elements and standards and a performance awards system as well. The performance appraisal system is not the major emphasis of this report. Rather, the report discusses an aspect of the Department's executive evaluation system that is unique among governmental agencies. In addition to receiving a performance rating based on accomplishments vis-a-vis performance standards, career executives are also rated on the risk/difficulty level of their positions. This rating is called a position coefficient rating.

Position Coefficient Development: The position coefficient concept was developed to resolve a problem related to the performance awards provision of the Civil Service Reform Act (CSRA) of 1978. That is, in establishing the performance awards program, the CSRA stipulated that performance bonuses be limited in any fiscal year to no more than 50 percent of the senior executive positions in an agency. Since these awards are to be based primarily on performance and since performance ratings are not subject to forced distribution, the 50 percent limit creates a problem for the award determination process. This is due to the fact that historically executives' ratings have clustered at the upper levels without a great deal of variance. If such rating patterns were to continue, additional criteria or procedures would be necessary to provide for finer discrimination amoung executives for selection of the award recipients.

The philosophical basis for the position coefficient concept is quite elementary and centers on the fact that some jobs are easier than others, some more controllable, some demand conservative execution, etc., and that these aspects of a position can be measured (i.e., rated). The position coefficient concept is designed to evaluate these differences between positions in terms of the probability for success in a position based on the risk/difficulty level of the position and to reward those executives who succeed in very difficult positions. The position coefficient rating is combined with the performance rating to determine award eligibility. Contrary to the performance evaluation system, the distribution of the position coefficient ratings is forced, providing more differentiation among executives. Two evaluation processes, one based on the evaluation of the position and the other on the performance of the executive in the position, produce ratings that are used in conjunction to determine the recipients of the performance awards.

The Rating Process: In accordance with the CSRA, the Department established Performance Review Boards (PRBs) to review performance evaluation ratings for executives and to make recommendations to the Secretary of Agriculture relative to final performance ratings, retention decisions, and performance award recipients. The PRBs are also responsible for the position coefficient ratings. The Department's PRBs, nine in number, were established along organizational lines to ensure that similar program areas or types of work would be included under the same PRB. This manner of composition of the PRBs is important to the position coefficient concept since it works best when the positions being rated do not involve significantly different executive functions.

The ratings are based on an evaluation of the career positions relative to four factors. Narratives by the supervisor, additional comments by the executive, and, where appropriate, comments by the second level supervisor are provided for each of the four selected factors. To facilitate the performance award determination process and to provide a means of combining the position coefficient and performance ratings, an award matrix was designed.

The position coefficient concept has been utilized in the Department for two years now and has met with mixed reactions from the Department's executives. In fact, there is a considerable amount of negativism toward the concept.

PAPER SESSION

Innovation and Alternative Selection Procedures

Chair: Jan Klein, CODESP

Discussant: Kenneth Krueger, Ability Information Systems

Personal Life History--Biodata Items as Suitable Alternative Selection Procedures: A Problem and Proposed Solution

Lawrence S. Kleiman, Virginia C. Falls, and Deborah L Wilcox
The University of Tennessee at Chattanooga

The Uniform Guidelines on Employee Selection Procedures (1978) specify that the validity of selection instruments is no longer sufficient in justifying their use if such instruments produce an adverse impact upon "protected group" members. Because traditional paper-and-pencil tests of cognitive ability frequently produce adverse impact, employers must search for alternatives. Biodata are potentially viable as an alternative; validity has been well documented (Cascio, 1976; Dunnette & Borman, 1979; Owens, 1976).

A concern exists regarding the fairness of biodata items, especially items dealing with personal life history. These items may perpetuate past discrimination by (1) penalizing protected group members who have been denied past opportunities to exhibit relevant (i.e., valid) behaviors, and (2) failing to account for recent changes in the behavioral patterns of many protected group members. A need exists to develop a type of biodata item which is at least as valid as the personal life history item, yet has less potential for unfairness. The "present life" (PL) item, assessing current or very recent behaviors of an individual, has been developed for this purpose. Items dealing with current behaviors would (1) minimize the effect of penalizing protected group members for previously denied opportunities, and (2) incorporate changes in behavior which may have occurred during an individual's recent past.

This study was concerned with determining the validity of these PL items. The Military Biographical Questionnaire (MBQ) was developed in an effort to reduce turnover in the Air National Guard. A 26-item personal life history questionnaire was developed from information based on behavioral patterns which differentiated between re-enlistees and non-re-enlistees. A PL item was written to correspond as closely as possible with each personal life history item, the major difference being the time orientation (past vs. present). Two forms of the MBQ were developed, each containing half PL and half personal life history items. The two forms were administered randomly among subjects from two squadrons in two cities. Re-enlistment intention served as the criterion measure.

Of the 26 pairs of item-criterion correlations, twelve PL and only six personal life history items were valid ($p<.05$). Within each pair of items, there were two significant differences ($p<.05$) between the coefficients, both favoring the PL items. A Wilcoxon Matched-Pairs Signed-Ranks test was conducted to take into account the cumulative differences between the coefficients over the entire set of 26 items. The items served as "subjects" (i.e., the unit of analyses); the nature of the item (personal life history or PL) served as the matched "conditions," and the item validity served as the dependent variable. The results indicate that the set of PL items possesses greater validity than the set of personal life history items ($z=2.17$, $p<.05$).

The major contribution of this study is that it offers an alternative type of biodata item which should be legally defensible, have less adverse impact and have at least equal validity. Further research is needed to extend the generalizability of these results to other settings and dependent variables. Research should also examine the actual adverse impact of each type of item, and should explore the dynamics of the success of biodata items.

The Development and Validation of a Self-Report Scored In-Basket Test in an Assessment Center Setting

Gerald A. Kesselman, Felix M. Lopez, and Felix E. Lopez,
Lopez Assessment Services, Inc.

In recent years, personnel selection specialists have expressed interest in using job simulations to predict managerial performance. The in-basket exercise is a situational test which presents the participant with a hypothetical work situation in which decisions must be made on a series of memos, latters, and other documents deposited as incoming mail. The general in-basket model includes three parts: a set of background materials, a set of problems, and appropriate feedback procedures.

In order to develop a frame of reference, the background materials typically include a description of the following: the company, organization charts, jobs, chief personalities, the calendar year, financial statements, and other pertinent information. The set of problems consists of documents and possibly phone calls or other oral messages to which the manager must respond. The third part of the model consists of the procedures used in which the participants record and explain their actions. The participant's product is either scored in some way, or reviewed with the participant by an interviewer or by a group of other participants. This aspect of the model has caused the most difficulty in using the in-basket as an assessment technique.

The findings regarding the validity of situational tests are mixed, and the authors hypothesize that this is due to the scoring formats currently used. This paper reports the results of an objectively based in-basket scoring key. The scoring format is a self-report questionnaire that was subsequently validated in an assessment center setting for key supervisory positions in an electric utility company.

The subjects were first-line supervisors in administrative and/or technical positions. Two types of job analyses were done: a job inventory questionnaire reflecting tasks and demands of the positions, and a threshold traits analysis used to determine the relevance, level, and practicality of 33 traits listed for acceptable job performance. Final job analysis results included a delineation of the traits and trait levels required, the weights for each relevant trait, and a listing of corresponding tasks and demands that supported the inclusion of each trait.

The job analysis indicated that similar abilities were required for both the administrative and technical supervisory positions. The in-basket test was designed to measure the traits of problem-solving, planning, and decision-making. A series of items based on work samples were constructed and tried out on a sample of company managers. Twenty-six items, some critical and others less important, were used. These items included acting on written materials and responding to a series of incoming phone calls. The time limit was two hours. An additional one hour was allowed to fill in an action report form indicating the priority and disposition of each item and giving reasons for the action taken.

Based on the try-out of the items, a list of 684 possible actions was developed and incorporated in an action report which is a self-report questionnaire. A panel of ten middle managers reviewed the action report to assign scoring weights to each action. Each action was assigned a scoring weight from 0 (totally inappropriate) to 3 (very appropriate) on three different dimensions: (1) problem-solving, (2) planning, and (3) decision-making. This panel also determined the priority of each item. Therefore, the candidate's total score on the in-basket test was a function of the appropriateness of the action taken weighted by the priority of the item.

The immediate supervisors of the subjects in the validation study completed two separate performance description reports measuring many of the same criteria. Form A, uniform for all jobs, was a graphic type performance description and measured performance on each of the 33 traits comprising the threshold traits analysis system. On Form B, a behavioral observation performance description report, each supervisor rated performance on a six-point scale on relevant job functions and traits determined by final job analysis results. The correlation between the two performance description forms averaged .70 across all the traits measured, indicating a moderately high degree of inter-form reliability for the criterion measures.

Eight criteria were utilized. Four were derived from each of the two performance description forms. The four separate criteria were: (1) comprehension and problem-solving, (2) planning and decision-making, (3) composite cognition (sum of ratings on perception, concentration, memory, comprehension, problem-solving, numerical computation, written expression, planning, and decision-making), and (4) overall job performance.

The three subscores on planning, problem-solving, and decision-making were highly intercorrelated (median r=.50) and were reflected quite well in the total score (median r of subscores with total score of .81; p<.01). The high intercorrelations among the subscores support previous in-basket research which suggests that the underlying ability measured by the in-basket is a single generalized trait (Lopez, 1966).

In a random sample of 30 cases, the split-half reliability coefficient was .83, supporting the hypothesis that an objective scoring key can yield in-basket scores that are quite reliable. The total in-basket score is positively and significantly correlated with six of the eight criteria. The median validity coefficient (uncorrected for range or criterion attenuation) is .26 (p<.05). The exercise correlated equally well with specific performance criteria and overall job performance ratings. The attenuated corrected correlations yielded a median validity coefficient of .31 (p<.01).

The authors suggested that future research should further refine the scoring methodology and include the use of advanced statistical weighting techniques. Future research should also investigate the extent to which deception may occur in this self-report scoring format, and recommend ways to minimize such deception.

Self-Assessment/Self-Selection: Implications from Research

Kelly F. Miller, Ph.D Candidate, University of Kansas

Self-assessment is the actual appraisal that candidates do of their own needs and skills. Self-selection is the candidates' use of the appraisal information, in conjunction with a realistic presentation of the organization and the taks of the available positions, to decide for themselves if they want to join the organization. Mr. Miller reviewed the available literature and research, and developed a list of practical suggestions for the use of self-assessment in personnel selection:

1. Beware of self-assessments that display the halo effect.

2. Structure the assessment to encourage the distinction of strengths and weaknesses.

3. Provide specific, readily observable information about the tasks for which skills are to be self-assessed.

4. Provide extensive information covering all facets of the task, and all contributing factors from every conceivable situation in which the task is to be done.

5. State any questions about skills in a value-free form such as found on a behaviorally anchored rating scale.

6. Provide applicants with training in the interpretation of the measurement scale and procedures.

7. Use self-assessment for screening applicants. Follow-up with tests for those who rate themselves: a) above the valid minimum requirements, or b) below the minimum, but have not recently used or been tested on the required skills.

Successful self-selection relies upon realistic information about the organization and the job. From the available literature, Ms. Miller compiled a list of when and how to best use such realism:

1. Realism is best used when the selection ratio is low.

2. A realistic job preview is more essential for the individual entering the organization than for someone making an internal transfer.

3. The information should be provided as early as possible in the applicant's choice process.

4. Realism is most effective during periods of high employment for the job type because applicants have a number of viable options. Negative information will be ignored if the primary objective is a job rather than job satisfaction.

5. There is an optimal level of realism that arouses vigilance and at the same time helps applicants build confidence in order to cope. This is similar to the expectancy theory notion that moderately challenging goals appear to be accomplishable.

6. Written or visual media promote realistic expectations.

7. Information must concern variables used by applicants in their choice process.

8. Realism is most effective in survival rates when the job involves a complex role.

Research on realistic job preview and self-selection is concerned with matching applicant needs and organizational characteristics. Correlational studies use job satisfaction, absenteeism, turnover, and secondarily performance as criteria.

SYMPOSIUM

Psychological Testing: Its Survival Problems

Chair: Clyde J. Lindley, Center for Psychological Service, Washington, D.C.

Presenters: Thelma Hunt, George Washington University
Bruce Davey, State of Connecticut
Sidney Teske, Hennepin County, Minnesota

The moderator, Clyde J. Lindley, set the stage for the symposium by quoting authoritative sources indicating that psychological testing still has to fight for its existence (theoretically and practically). He made particular reference to the recent National Research Council, National Academy of Science study of testing. A statement from the Academy's report which he quoted is particularly pertinent to IPMAAC concerns.

> Advocates of testing consider it the best available means of impartial selection based on ability; many are, in addition, enthusiastic about the value of tests in revealing undiscovered talent and extol their contribution to increased efficiency and accountability in a variety of educational and employment settings. Critics of testing have found the negative effects of testing more compelling. They claim that tests measure too little too narrowly. And some spokesmen for minority interests have attacked standardized tests as artificial barriers to social equality and economic opportunity.
>
> Both high expectations and serious complaints have focused public attention on the underlying questions of what tests actually measure and the meaning to be attached to test scores. The increasing interest of courts, legislatures, and governmental agencies in the way tests are used in selection systems has added a significant new dimension to these questions.

The moderator also clarified the meaning of the symposium considerations by defining psychological testing to include evaluations of abilities, aptitudes, achievements, interests, skills, motor abilities, physical fitness (as strength and agility); as well as personality and psychological fitness characteristics (often considered to be the scope of "psychological" tests).

Lindley appropriately brought in the public's interest by stating that "the public's interest in psychological testing as a personnel problem results from the fact that almost everyone is affected directly or indirectly by selection or promotion decisions in the work place which makes use of some type of test. Those affected want to be certain that the test instruments are fair and socially non-discriminatory."

As a symposium member, Thelma Hunt dealt in a general way with the survival problems for psychological testing, orienting her presentation toward personnel uses of tests. The concerns of her remarks were mainly directed toward the "threats" to psychological testing. Her take-off theme was: If there were no threats to psychological tests, there would be no survival problems.

Some of the threats to the use of tests in personnel selection relate to attitudinal line-ups of potential users and advocates. How do you as a test expert or advocate react on the firing line? Can you go along with the "avant-garde," and accept Schmidt and Hunter's theory of generalizability of employment tests? Do you just toe-the-line on what has been set down and supported by experience? Do you take the laissez-faire approach of "Don't rock the boat"? Does everything have to be integrated before you can move? Do you give up - we can't win with psychological testing as a selection technique? Think about your attitude. It may have some surprising effects.

Dr. Hunt further commented on specific threats. Her major contribution had to do with threats that relate to general ability testing. General ability tests have had various labels: intelligence tests, I.Q. tests, mental ability tests, general aptitude tests, tests of learning potential, etc. Many have not had identifying labels suggestive of their general ability testing nature such as the court famous Wonderlic test. Practically all verbal pencil and paper employment and promotion tests involve a strong component of general ability testing. The challenges against testing (particularly in the courts) have mainly been against tests with strong general ability components.

Many accepted tests used in the personnel field are <u>specialized</u> general ability tests; that is, tests measuring general abilities couched in job relevant language and situationally related context. Examples are wr tten tests for entry-level clerical workers, tests for selection and p1)motion of police officers and fire fighters, and tests for law school applicants. Many personnel tests have become less and less specifically oriented and more general ability oriented. This trend is in recognition that entry and promotion tests should emphasize general ability potentials, with a subsequent emphasis on job training.

Dr. Hunt then spoke of some, perhaps more often spoken of threats to psychological tests. On the problem of differential distribution of test scores, she commented that one of the significant problems with which we have to deal is the assumption, on the part of test challengers, that a test producing a differential distribution of applicant scores (unfavorable for the disadvantaged applicant group versus the perceived advantaged group) is by that fact challengeable with respect to the test's validity. But a test that is valid for one group is nearly always valid for the other group. High score applicants do better on the job than low score applicants in all groups. Nevertheless, differential distributions of scores need to be studied in relation to job performance so that a given job performance can be equated to the proper score (predictive of that performance) in each distributi^n of scores.

Among other things, she discussed job performance evaluation as a threat factor. While there may be other less frequent uses, psychological tests in personnel work are mostly to be considered in employee selection or promotion. So, there is the necessity of adequate job performance evaluations for validating psychological tests used for selection and promotion. It is unfortunate that the greater blame in interpreting low validity findings, on the basis of which tests have been rejected, has more often been placed on the test without critical evaluations of the criterion measure (job performance).

Sidney Teske discussed three threats to psychological tests:

1. The external threat from what he called the public,

2. The threat from within the administrative and management ranks of our organizations, and

3. The threat we pose to ourselves.

He sees adverse public opinion toward psychological tests, in considerable measure, as a generalization from many things often not directly related to tests. In public employment (IPMAAC's sphere of operation), the public typically hears about government workers because of some scandal involving a public employee who has misused public time or money, or because an EEO suit or charge is made. Generalization may go to adverse opinion toward the selection procedures that put these employees in their jobs. Teske sees a need for sympathizing with job applicants whose adverse attitudes, from standing in a long applicant line for various aspects of the selection process, rub off on psychological tests which they may have to take. According to Teske "we must be strong advocates for the responsible treatment of applicants when we come in contact with them from the point of recruitment through testing and their 'shelf' life on an eligibility list."

Teske thinks assessment professionals should be a bit more aggressive than in the past in relation to decision-making managers and administrators. He said assessment professionals must be willing to use more definite language when talking with managers, be committed to a strong testing program and defend the value of the ranking process. He proposed that we avoid the use of language which leads to insecurity simply because we want to hedge and avoid making a decision which has a possibility of failure. Stick your neck out and work hard at keeping your neck by building good tests

Teske's third point addressed the important question of the psychological testing expert's devotion to sound testing principles and social goals. To quote: "Should I agree that sound testing principles can safely be voided in favor of a social goal that I might also support? I would like to propose that the scientist should not do so. We have a big enough task on our hands trying to maintain the use of good testing techniques without confusing the decision makers about the best approach. I believe that when these questions arise, my role should be that of a professional defending the quality of an objective selection process. I also believe that the social argument must be made, but not by the test professional."

In his contribution to the symposium, Bruce Davey first spoke of the popular mistrust of tests. His interesting presentation can be conveyed best by using his own words:

> One overall impression I have is that there sure are a lot of people out there who are unhappy about tests. Tests are about as popular as root canal work. There are a lot of people out there who might bomb this building if they knew there were so many testers under one roof.
>
> Of course, we shouldn't expect our profession to be too popular. People don't like to be assessed, they don't like to be measured as specifically as a test measures them. From school age on, people have been threatened by tests, and everyone has been wounded one or more times, or maybe repeatedly, by tests. Their egos get bruised and they don't forget. I've done a fair amount of research on self-assessment, and one interesting thing I've observed is that most people consider themselves to be highly competent compared to their peers. In fact, the average person rates himself at around the 85th or 90th percentile--and that's good, because a good self-image is good for one's mental health. But, tests threaten our inflated egos because when people compete on a test, half of them score below average, and most people consider being below average akin to lying in the gutter. And so the tests are attacked. And the more valid you say your test is, the more threatened people feel by it, and the harder they're apt to fight it. I suspect that some of the pressure on testing today is a result of the collective human psyche fighting back against a perceived enemy. I wonder if testing would be more popular if the test results were scaled to match up more favorably to people's ego needs. Perhaps we shouldn't report exact numbers--perhaps we should only report positive-sounding descriptive labels. For example, scores from 90 to 100 might be described as "Fantastic"; scores from 80 to 90 might be labelled "Exceptional"; scores from 70 to 79 would be called "Very Good"; scores from 60 to 69 would be labelled as "Not bad at all"; and under 60 would simply be labelled as "Promising."

Davey pointed out in several ways that tests should not be the focal point of the issues involved. Tests are not creating adverse impact out of thin air--they are discovering adverse impact, and it is impact that occurred long before the test was given. The test is uncovering a social problem which a lot of people do not want to face. The problem is that some groups have not been given an equal piece of the pie before, and so they are at a competitive disadvantage as a group when they compete with the advantaged majority on standardized tests.

Davey took a swipe at some official capacities by saying that there are a lot of legislators and government officials who seem to have decided that focusing on tests and getting them removed will solve the problem. But of course it will not. What replacements are available for tests? Interviews? A lottery? Patronage? In any of these cases, the research that has been done indicates that tests are more valid than interviews, and certainly more valid than a patronage system or lottery.

Davey ended his remarks on a very optimistic note: "I see a lot of new optimism to counter the old problems, and I think that bodes well for the survival of testing."

Western Region Intergovernmental Personnel Assessment Council (WPIPAC) INVITED SPEAKER

The Cutting Edge of Selection Developments

Dr. Robert L. Ebel, Michigan State University

There are two kinds of people in the field of measurement: theoretical psychometrists who deal with abstract concepts, and technicians who put the theories into practice. Dr. Ebel addressed some of the problems that need to be solved in the practice of testing and measurement. In considering the topic of the "cutting edge," he expressed the idea that excellence is often sacrificed for novelty. Some of the best methods and biggest discoveries have been developed long ago, and now we are heading towards more remote topics rather than dealing with basic issues.

One basic issue deals with problems associated with the determination of passing scores. A large distinction (e.g., qualified vs. unqualified, pass or fail) is made between people who may have very small differences in test scores; yet, there is no way to be totally objective in setting a passing point. There is always room for argument. The examiner makes the decision of how much the examinee should know. Because different examiners have different standards (the Achilles' heel of absolute measurement), in most cases, normative standards of measurement are used. The idea is to select those that know the most because it is impossible to determine an absolute amount that one must know in order to be qualified.

A second issue involves the use of written tests to assess cognitive ability. Today, there is an emphasis on other characteristics such as personality or motivation, however, these characteristics cannot be easily measued by paper and pencil tests. Dr. Ebel stated that knowledge is a structure that the individual develops in order to integrate and understand the relationship between concepts. Any cognitive ability depends solely on the person's knowledge. Knowledge about how to do the job is sufficient to have the ability to perform the job. Skill at performing the job is developed with practice and repetition. In personnel measurement, KSA's are treated as if they were distinct, but they all rest on knowledge. One must have the relevant knowledge to have the ability, and one must practice the ability to develop the skill.

Another issue deals with the use of multiple-choice and true/false items. Dr. Ebel stated that since decision-making and the ability to choose are essential cognitive abilities, which format could best assess those abilities? Simple propositions can be used to test knowledge merely by changing a key word or phrase. Dr. Ebel developed the alternate-choice items in which the examinee is presented with two choices and must select the word or phrase that best completes the statement. In a study he conducted on a class of his students, Dr. Ebel found an average reliability coefficient of .67 for alternate choice tests, as compared to .42 or .43 for traditional true/false tests. Dr. Ebel stated that testers combine too much information in one question in order to get four responses for traditional multiple-choice questions. More information can be obtained by asking separate questions with separate scorable responses. Simple test items can test whether the examinee has the necessary concepts and understands the relationship between them in order to do the job. Complex test items lower the reliability of test scores because they do not discriminate as well between examinees.

Another issue is that of establishing and measuring the validity of tests. Validity originally meant the extent to which the test measures what it is supposed to measure. Now, there are various types of validity: content, predictive, concurrent, and construct. Dr. Ebel stated that construct validity was never a very good idea, and cited Cronbach's contention that construct validity is a better topic for the seminar room than for the public arena. Construct validity is appropriate if one is discussing personality constructs. It requires the validation of predictions made from theory, and theories about personality are weak and imprecise. Dr. Ebel proposed the use of operational validity: does the test do the job? In order to establish operational validity, one cannot rest an argument solely on empirical evidence, but must rely on rational inferences and logical arguments. This will not yield a validity coefficient, but one can argue rationally that one test is more valid than another. With construct validity one can conclude only that the two tests measure different constructs.

The key to a successful test is good test items, which depend on how well the items are written. How well the items are written depends on two things: what kind of knowledge base the writer has and how literate the writer is. The test writer must be able to express ideas correctly and concisely so that the items will work effectively. This is the area of the field that needs the most work.

STUDENT AWARD

Chair: Glenn G. McClung
IPMAAC President

Jerry Thompson, Member
IPMAAC Public Affairs Committee

Recipient: Kenneth Pearlman
The George Washington
University

The International Personnel Management Association Assessment Council sponsored for the first time a Student Paper Competition Award. This award was presented at the Annual Conference on Public Personnel Assessment, June 6-10, 1982 in Minneapolis, Minnesota.

Ten students submitted abstracts on topics including "Reasonable Accommodation: Key to Successful Employment of the Disabled in the Public Sector," "Comparison of Four Approaches to the Evaluation of Job Applicant Training and Work Experience," "Officer Development Program: A Study of Stress Management," etc. These students were in attendance at the following universities: California State University at Long Beach, Southern Illinois-Carbondale, University of Kansas, University of Akron, The George Washington University, Memphis State, University of Illinois, University of Connecticut, Penn State University, and University of Missouri.

These abstracts were sent to committee members for ranking, and the authors of the top five abstracts were invited to submit papers for evaluation for the final award. The Student Award was presented to Kenneth Pearlman who, at the time, was a Ph.D. candidate at The George Washington University. Dr. Pearlman is employed at the U. S. Office of Personnel Management. He received a $300 award in recognition of his contribution to the field of public personnel assessment.

The Bayesian Approach to Validity Generalization: A Systematic Examination of the Robustness of Procedures and Conclusions

Kenneth Pearlman

For over 60 years, the validity coefficient (i.e., the correlation between individuals' scores on a written test or other selection procedure and a measure of their performance on the job or in training) has been the most basic and widely used tool for assessing the predictive value of a selection procedure. By the mid-1940's and early 1950's, after fairly large amounts of validity data for different types of tests and jobs had accumulated in the research literature, personnel selection psychologists were becoming increasingly interested in assessing the consistency of their validation results and the knowledge gained through such efforts. Ghiselli pioneered this movement by reviewing and integrating the results of hundreds of published and unpublished studies of the validity of different types of tests for a large number of occupations. His hope was that "from the generalizations about validity given by an overview, inferences could be made about the validity of specific tests for specific jobs, which possibly would be more generally applicable than the results of any one investigation."

Ghiselli was thus among the first to articulate the basic concept of validity generalization--the extrapolation of empirical validity results to new jobs or settings for which empirical data are unavailable--as well as to delineate some of its potential benefits to the field. However, Ghiselli's hopes of finding readily generalizable test validities were soon dampened by the results of his research, which revealed generally unimpressive levels of average test validity and showed that different types of tests exhibited no typical level of prediction for any given occupation. Rather, there was a high degree of variability in the validity results from study to study, even when the tests and jobs studied appeared to be similar or essentially identical.

The effect of Ghiselli's studies was to reinforce, and for many psychologists, to confirm the widely held belief in personnel/industrial psychology that test validity is highly situationally specific (i.e., that it is highly subject to the influence of situational or between-job moderator effects). As a result, it was concluded that empirical validation was required in each situation and that validity generalization was essentially impossible. Although some psychologists recognized this state of affairs to be perhaps the most serious shortcoming in modern selection psychology, most resigned themselves to the harsh "facts" of situational specificity and the inability to generalizr validity.

The most important consequence of this apparent inability is that it precludes development of the general principles and theories of selection that could take the field beyond a mere technology to the status of a science. The first step in the development of general principles and theories in this (or any other) area is the establishment of stable patterns of relationships among basic variables. In order to establish such relationships, situational specificity must first be demonstrated to be essentially false or have limited impact. If situational specificity can be rejected, then relationships between various constructs (e.g., verbal or quantitative ability) and specified kinds of performances are, by implication, invariant at the population level. The emergent pattern of such population parameters and their interrelationships would thus form the basis for a science of personnel selection.

The practical implications of validity generalization are also considerable. There are many selection situations in which it is technically infeasible to conduct an empirical validation study. Even when technically feasible, such s.udies are often prohibitively expensive. The ability to generalize the results of previous validation studies of particular types of tests across settings would obviate the necessity for carrying out empirical validation in each new setting. It would thus circumvent the problem of technical infeasibility, as well as lead to potentially large dollar savings by eliminating the need for many criterion-related validity studies.

In 1977, Schmidt and Hunter introduced a new approach and potential solution to the problem of validity generalization based on their hypothesis that most of the observed variability in validity outcomes from study to study might be due to various statistical artifacts rather than (as had long been believed) true factor structure differences among apparently similar jobs (i.e., true situational specificity). The Schmidt-Hunter procedure is based on correcting the observed mean and variance of a distribution of empirical validity coefficients for such artifacts (e.g., sampling error, and between-study differences in test and criterion reliability and range restriction) in order to produce a distribution of true (i.e., unrestricted and unattenuated) validities. This corrected distribution is then viewed as a Bayesian prior distribution whose properties (i.e., mean, standard deviation, and confidence limits) quantify the degree of validity generalization characteristic of the given test-job (or job family) combination represented by the distribution.

Since the introduction of a Bayesian approach to assessing validity generalization. the viability and practical applicability of this approach has been demonstrated in some 20 different validity generalization studies conducted by Schmidt, Hunter, and colleagues and by others. These studies found that just three or four of seven potential statistical artifacts accounted for an average of about 70% of the observed variation in distributions of empirical validity coefficients; and, that a conclusion of validity generalization for the predictors in question (mostly traditional cognitive ability measures, but in some cases such predictors as performance and psychomotor tests, biodata, and college grade point average) was supportable in over 85% of all such distributions. This conclusion held for individual jobs, for relatively task-homogeneous job families, and for larger populations of jobs spanning considerable portions of the occupational spectrum. These studies, taken collectively, have cast serious doubt on the doctrine of the situational specificity of employment test validities and have demonstrated test validity to be far more generalizable than ever believed. That is, they have shown that when subjected to statistical corrections and procedures appropriate to the analysis of results cumulated across individual studies (i.e., meta-analytic methods), selection procedure validities are in fact considerably higher and considerably less variable than had been apparent in the past.

Notwithstanding these positive results, there have as yet been only a few beginning attempts to assess the full power and robustness of the Bayesian approach to validity generalization. A number of important methodological and substantive issues have been raised by the introduction of different possible statistical procedures and data grouping methods within the same basic validity generalization strategy. In particular, four unresolved (or only partially resolved) issues may be defined from the current state-of-the-art in validity generalization research. These center broadly around the question of the robustness of the phenomenon of validity generalization relative to different possible statistical and conceptual treatments of validity data. (The term "robustness" is used here in the traditional statistical sense of lack of susceptiblity to violations of assumptions and/or departures from optimal procedures.)

The purpose of this study was to systematically investigate each of these issues, which was the next logical step in the programmatic research on Bayesian validity generalization methods. The general hypothesei of this research was that validity genrealization is a very broad and robus phenomenon. It was hypothesized that conclusions regarding validity generalization would be unaffected by: (1) potential moderator effects across very different types of jobs; (2) different possible bases for combining validity data for different jobs; (3) fairly extreme violations of assumptions; or (4) procedural variations.

The first three of these hypotheses were tested using some 2,400 validity coefficients (r's) drawn from 500 independent, large-sample validation studies of 61 enlisted occupations in the U.S. Navy. The 61 jobs vary widely in their task makeup, occupational level, and skill and ability requirements. The predictors in each study were the subtests of the Navy's Basic Test Battery (BTB), used for the selection and classification of Navy enlisted personnel from World War II through the mid-1970's. The BTB consists of six subtests, three of which (GCT, ARI, and CLER) are measures of traditional ability constructs (verbal reasoning, arithmetic reasoning, and perceptual speed, respectively) and three of which (MECH, SP, and ETST) are to varying degrees measures of rather specific knowledge and information (mechanical knowledge, shop information, and electronics information, respectively). The criterion in all studies was success in training, as indexed by individuals' final grade in each occupation's training program--a composite of various achievement and performance measures of the knowledges and skills learning in training. These data had never been previously extracted and used for any type of cumulative analysis.

To address the first two hypotheses--the effects on validity of potential between-job differences and of alternate job grouping strategies--the validities of each BTB subtest were pooled both across all jobs and within job families formed according to each of four different job analytic bases most commonly used in the formation of job families (molecular work content, worker-oriented job content, human attribute requirements, and the broad content structure or overall nature of the job). To the extent possible, more than one specific job grouping system was used to represent each of these four substantive categories of job grouping (a total of 15 specific job family systems were ultimately used). This was to allow for later comparison of results among alternate systems within the same category and to increase the reliability of results obtained for each category by averaging results across systems of each type. In addition, a number of random job groupings, each consisting of different numbers of job groups, were formed such that there was a system of randomly grouped jobs having the same number of groups as each of the substantively based systems. This was to permit later comparison of results on an equal group-size basis between randomly and substantively grouped jobs, which would allow for estimation of the effects on validity of different substantive systems, independent of the effects due purely to the number of groups in the system.

The pooling of validities for each BTB subtest across all jobs and into the groups defined by the various substantive systems produced a large number of validity distributions (test - job group combinations) for subsequent analysis. The procedure used was a variation of the Schmidt-Hunter procedure, a simplified and relatively assumption-free adaptation of which was made possible by the quality of the Navy data, which allowed for very precise estimation of the effects of artifacts on validity.

The third hypothesis--examining the effects of violations of underlying assumptions of the validity generalization procedure used--was also tested on the NAVY BTB data base described above, using a very powerful method of examining such effects: by assuming there is no source of artifactual variance operating in these validity distributions other than ordinary sampling error. That is, the distributions were analyzed by correcting the variance for just this one artifact to see whether validity generalization conclusions would change from those obtained in the previous analyses, despite the fact that several statistical artifacts (e.g., between-study criterion reliability and range restriction differences) known to be operating in these data were not corrected for.

The fourth hypothesis--examining the effects of using alternate validity generalization procedures--was tested with a large-scale empirical comparison among four different validity generalization procedures advanced to date: three developed by Schmidt and Hunter and one by Callender and Osburn. These four procedures (each of which is based on slightly different underlying conceptual and statistical assumptions) were used to analyze 56 validity distributions gathered for an earlier validity generalization study of clerical occupations involving ten different test types (ability constructs), five clerical job families, and two classes of criteria--job proficiency and training success.

The results of all the above analyses provided strong support for all hypotheses. Specifically, it was found that:

1. Test validities were highly generalizable not only for individual jobs and relatively homogeneous job families, but across a wide range of jobs representing most of the occupational spectrum. These findings mean that whatever differences in tasks, behaviors, job complexity and responsibility, job-related worker abilities, or situational characteristics exist among these jobs are not sufficiently large to produce a significant moderating effect on validity; that is, to change the conclusion that each predictor is valid for every job.

2. Differences in usefulness for selection applications among the four major bases for grouping jobs were negligible. Although all grouping systems showed some degree of usefulness, it was found that on the average roughly half the gain in effectiveness from grouping jobs (as opposed to treating all jobs as a single, undifferentiated group) was a function of the number of jobs in the system rather than any inherent characteristics of the grouping systems themselves. These findings support the conclusion that simple, rational groupings are as useful for selection applications as groupings normally derived by more complex, time-consuming, and expensive methods.

3. Validity generalization conclusions were supportable on the basis of a simplified validity generalization procedure that was free of most of the assumptions usually required of such procedures. Based on the analysis of 750 different validity distributions, it was found that not only were conclusions essentially unchanged when only a single artifact--sampling error--was corrected for, but the loss in accuracy of results was very small despite the extreme violation of procedural assumptions usually made in validity generalization research.

4. Alternate procedures for estimating the critical parameters in validity generalization research produced essentially identical results and conclusions when applied to the same empirical data set. These findings indicate that any of the four current procedures may be appropriately applied in practice.

In evaluating the results of the various analyses carried out as part of this research, it is important to bear in mind the extraordinary empirical foundations on which they are based. The Navy data consisted of over 2,400 validity coefficients representing a total sample base of over 1.75 million. The clerical data consisted of over 2,650 coefficients representing a total sample size of over 375,000. These represent two of the three largest compilations of validity data ever brought together for cumulative analysis. The tests and jobs included in both data sets represent highly typical selection situations occurring in all types of organizations in both public and private sectors. Thus, the conclusions reached in this study can be considered highly reliable and generally applicable.

Results of this study confirmed a number of findings from earlier validity generalization research, extended some of these findings in significant ways, and addressed several major issues never previously explored. For example, results confirmed earlier findings that the validity of typical selection tests is higher and more generalizable than had ever been believed (with correspondingly greater gains in overall workforce productivity available through the use of such selection procedures than had been previously realized); variation in validity across a very wide range of jobs is sufficiently small to allow conclusions of validity generalization across the spectrum of occupations for which tests are most commonly used in selection; and validity generalization conclusions are supportable on the basis of corrections for sampling error alone. For each of these findings, the range of occupations, variety of job groupings and predictor types (knowledge/ information measures as well as ability measures), and nature of the validity data (both quantity and quality) on which results were based represented notable extensions of prior research on these issues. The comparison among alternate validity generalization procedures was the first large-scale empirical analysis of its kind, and the results conclusively demonstrated the trivial effects of method differences on research conclusions.

Perhaps the most interesting and unique facet of this study was the systematic assessment of potential within-and between-group moderators of validity for 15 different job grouping systems representing each of the four major bases for job family development in common use. This analysis represented the first actual quantification of the usefulness of alternate job grouping systems for selection purposes in either the validity generalization or job families research literature. Results explicitly ruled out the possibility of the most common types of job descriptors having a substantial moderating effect on validity, and showed that all major types of grouping systems have about equal (though very modest) effectiveness for selection applications.

Taken together, these findings strongly support the general hypothesis that validity generalization is a very broad and robust phenomenon As a result, they further solidify the groundwork for future theory development concerning trait-performance relationships in the world of work, a prerequisite to which is the establishment of the types of stable attribute-performance relationships observed in these analyses. In terms of practical implications, these results should pave the way for widespread application of Bayesian validity generalization procedures in the development, refinement, documentation, and evaluation of employee selection programs for both individual and grouped jobs.

From a broader perspective, this study can be viewed as bringing to a close what might be considered the first phase of a long-range research program designed to construct a body of general principles and theories out of the widely scattered and seemingly chaotic empirical results of over 60 years of personnel research. This phase has necessarily focused on dispelling the many myths within which the field had become encrusted. Foremost among these myths was that of the situational specificity of test validity and its concomitant misbegotten prescriptions for practice in the field, such as the need for local, in situ validation of all selection procedures, the need for the most detailed and comprehensive job analysis possible, and the injunction against combining any but the most demonstrably identical jobs together for validation studies or common selection procedure use.

There would appear to be little more that could be done to lay such ghosts to rest beyond what has been demonstrated in this study, in conjunction with the several previous large-sample validity generalization studies involving wide ranges of tests and jobs. It is now time to begin the second phase of programmatic, theoretical research in this area, taking full advantage of the powerful meta-analytic procedures currently available, of which the Bayesian approach to validity generalization is one example. The first step in this process, which is already underway, should be to complete the meta-analytic synthesis of presently existing train-performance data, while filling in necessary gaps through large-scale or cooperative research studies. The combination of improved data reporting practices, the meta-analytic integration of existing data, and new research and data integration where needed should lead to dramatic progress in developing a science of personnel selection in the years ahead.

GREAT LAKES ASSESSMENT COUNCIL PANEL: Selection In The Private Sector

Chair: John D. Sprague, City of Minneapolis

Presenters: Richard Arvey, University of Houston
Gail Drauden, Honeywell, Inc.
Norman Peterson, Personnel Decisions Research Institute
Gary Stormoen, Minnegasco

(No written material was available for the first three panelists; the tape was unintelligible)

Gary Stormoen identified characteristics that are particular to the private sector, as seen from his experience at the Minnesota Gas Company. (He pointed out that this company is by no means typical of all private enterprise.)

The selection system at Minnegasco was developed by Cliff Jergenson, who had a penchant for the collection of data. In the company archives, there often exists as many as ten or fifteen test scores for employees between the years 1950 and 1973. Three or four scores exist for employees subsequent to 1973. Also, criteria data were collected on each individual including tenure, salary progression, absenteeism, productivity, and job performance. Stormoen stressed that this wealth of data is atypical of the private sector.

Minnegasco primarily makes use of differential aptitude tests measuring such things as verbal reasoning, abstract reasoning, and mechanical comprehension. In spite of the problems that exist with the use of these types of tests (such as litigation and adverse impact), they will probably continue to be used due to certain organizational constraints. Some of these constraints are:

1. In private enterprise, the emphasis is on the person's career rather than on the specific job. Professional positions within the company are not static but are constantly changing. Among union ranks, all promotions and transfers are based primarily on seniority. It would be nearly impossible to develop content valid tests for each of these positions.

2. Tests are administered before hiring, and then only two or three times again during the person's employment with the company. Neither employees or management would desire that all candidates undergo testing for each promotion.

3. A great deal of training takes place, both formal and on the job. Among union employees, aptitude tests are used for initial selection, and then extensive training is given. This has created such a homogeneous group of employees that seniority is the most appropriate method of selection for promotion. (Test batteries have been validated against basic management criteria for all positions in union personnel.)

Tests are primarily used to provide as much discriminatory power as is feasible, and to use the information provided by the tests to make the best selection decision possible. Applicants are never strictly rank ordered, and there are no hard and fast cut-off points. Scores for candidates competing for a certain position may be differentially weighted. Protected class members are not treated identically within this structured process.

Certain problems exist with this system. These include:

1. Thirty-five years of testing have produced a tremendous restriction of range. Concurrent validation studies cannot be conducted because virtually no variance exists.

2. The organizational climate in the private sector can change quickly, thus, the types and methods of selection decisions can change quickly.

3. Merit system regulations and laws do not exist in the private sector. The testing program rests upon its acceptance by management. Nothing compels them to make use of testing procedures in selection.

PAPER SESSION

New Developments in Selection

Chair: Susan Biesele
Salt Lake County, Utah

Discussant: Richard C. Joines
Management and Personnel
Systems, Inc.

Validation of a Multi-Jurisdictional Test for Employment Service Interviewers

Steven S. Nettles and Michael Rosenfeld
Educational Testing Service

This paper focused on the development and validation of selection procedures for Employment Service Interviewers for use in all 50 states in the U.S. Although the total study also included work on an oral examination, this paper was limited to a criterion validation study of a developed written test. The study began in 1974 and was funded by the U.S. Department of Labor.

The first phase of the study entailed conducting an extensive job analysis involving 1600 task inventories administered in 17 states to determine whether a "nationwide" position exists. Factor analysis produced seven job dimensions, and it was determined that three of these were performed by most employment service interviewers across the nation. These three were: conducting personal interviews with applicants; selecting, referring, and placing applicants; and job order-taking.

In phase two, a training program was developed, administered and evaluated to familiarize state merit system and employment service personnel with the test that was developed and the criterion procedures to be used in the study.

Phase three involved a criterion validation study of the written selection test conducted in four states. A job analysis task inventory was administered to 316 job incumbents to determine that all the incumbents being tested actually performed the job for which the test was developed. The final N was 262, including those participants who performed at least half the tasks in the factor analytically determined job dimensions. The subjects in this phase of the study also had to be engaged in developing and maintaining professional relationships with employers and other community groups.

Three criterion measures were used in the study: (1) a multiple choice work knowledge test assessing knowledges necessary to perform the job tasks; (2) a placement simulation exercise consisting of three separately timed parts based on a fictitious company trying to staff a new operation in a rural location; and (3) supervisory ratings on 36 items rated on a one to four scale.

High reliabilities were reported for the written selection test being validated, and for the work knowledge test and the placement simulation exercise. Data for determining the reliability of the supervisory ratings were not available. There was a significant race effect in the supervisory ratings. White interviewers were rated higher than black interviewers regardless of the race of the supervisor. A moderate correlation coefficient (.53) was reported between the work knowledge test and the placement simulation exercise. Low correlations were found between the supervisory ratings and the other criterion measures.

Substantive evidence of validity of the selection test was found for black and white and male and female job incumbents against the two objective measures (work knowledge and placement simulation performance).

In summary, Nettles pointed out that this study illustrates the value of consortia activities. He considered it a cost-effective method for consideration of the use of tests. It also showed that a criterion-related validation study can be conducted with reasonable resources.

Job Simulation Test for English Language Proficiency for Foreign Trained Nurses

Barbara A. Showers, State of Wisconsin

This study was carried out in the State of Wisconsin, and the introduction set the stage for her presentation.

One of the requirements for licensure as a registered nurse in the State of Wisconsin is the ability to speak and understand English in the health care setting. Existing tests of English language proficiency were primarily written examinations testing aural comprehension on general topics. No test appeared to be available which tested both aural comprehension and ability to communicate verbally with others in the health care setting.

A pilot project was undertaken to identify the critical communication tasks of a registered nurse and to develop an examination which would adequately assess them. A job simulation approach was selected for its potentially greater validity. It was felt that the critical incidents could be realistically portrayed by means of video tapes. Candidates could then take the test by viewing the video tapes and responding into a cassette tape recorder.

Critical incidents in the job relating to the topic under consideration were identified through a panel of registered nurses, nursing supervisors, and nursing educators. Five major types of English language interaction in the health care setting were identified. They were:

1. Taking instructions over the phone and reporting them accurately to the next nurse coming on duty.

2. Giving descriptive information over the phone, e.g., reporting a serious change in a patient's condition.

3. Taking instrıctions from another person in an emergency situation, and repeating the instructions.

4. Taking a patient history. Obtaining the pertinent information from a rambling monologue, recording it on the patient history form, and reading it back.

5. Explaining a medical test procedure to a patient so that the patient understands the test and the role the patient plays in the test.

The specific incidents identified for each type of interaction formed the basis for the scripts which were written.

Choice of Test Method. A method was chosen which would simulate the critical incidents as closely as possible, allow for an oral response from the candidate, and be as standardized as possible.

Video taping allowed for realistic, standardized situations which could be repeated for every candidate. Audio taped responses allowed the candidate to respond to the situations and be recorded for future evaluation.

Script Writing. The panel of nurses and nurse educators who identified the critical incidents provided the detailed situations which were developed for the scripts for the video tapes. Five video tapes were produced. These constitute the test as developed and validated.

Scoring criteria were developed mainly from ideas gained in the critical incident identification. Three sets of 9-point bench mark scales were developed for the dimensions of vocabulary, articulation, and comprehension. Each concept was defined and bench marks established for "more than acceptable," "acceptable," and "less than acceptable." The same scales were used for all scripts.

Reliability and validity studies of the test were reported on 19 registered nurses in Madison and Milwaukee, Wisconsin health care facilities: 12 native English speaking; 7 foreign trained speakers of English.

Reliability was reported in terms of consistency in ratings by different raters. As summarized by the author: "The reliability results as a whole appear to indicate that the raters could distinguish consistently between native and non-native speakers of English, and between individual non-native speakers. Given the small sample size of the non-native group, further data are needed to verify reliabilities. However, the current results do appear to be encouraging for future test use with foreign trained nurses."

Validity. Content and construct validity were discussed, and it was proposed that video taped job simulation tests provided a more direct assessment of functional on-the-job language requirements than existing written and listening comprehension tests which do not assess health care related language skills, or spoken English in the health care setting.

Criterion validity data as presented are more variable and less convincing.

Showers presented a very interesting study with innovations in methodology and an attack on an important type of communication problem. The only negative criticism, which is not really a criticism, is the pointing out of the need for a confirming study on a larger numbe. of cases.

Scale Values for 111 Words

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

The purposes of this study were:

1. to determine, on a scale of 1-10, the scale values of words that are commonly used in performance evaluation, rating of job applicants, job analysis, and other personnel management applications; and,

2. to determine if there are differences in the way various groups perceive these words.

For example, is being "satisfactory" better or worse than being "average"? Are they even on the same behavioral dimension? Is one person's perception of what is "unsatisfa-tory" likely to be the same as another's? Is "average" a positive, negative, or neutral term? This study investigated these and other questions relating to potential discrepancies between the intended use of a word and its actual perception by another person. The objective was to produce data which would enable people who construct rating scales to choose scale anchor points which produce meaningful results.

Respondents were asked to rate words on a scale of one to ten. Part I contained 56 "quality" type words (very good, satisfactory, needs improvement, etc.). For this type of word, a rating of "one" indicated this was not a very good thing to say about someone. A rating of "ten" indicated a very good thing to say. Parts II and III dealt with "importance" words and "frequency" words. For the thirty "importance" words, ratings ranged from one (not very important) to ten (very important). For the twenty-five "frequency" words, ratings ranged from one (not very often) to ten (very often).

Responses were solicited from as wide a range of people as possible. Five hundred forms were distributed with a return from 352 people. After screening for evidences of carelessness and invalidity, 265 usable returns remained for study. Sufficient numbers of respondents were obtained to permit comparison of the following groups: (1) men and women; (2) older and younger people; (3) government and nongovernment employees; (4) more educated and less educated people; (5) personnelists and nonpersonnelists; (6) different types of personnelists; (7) clerical, managerial, technical, and administrative employees.

Means, standard deviations, and lower and upper 5% confidence intervals were computed for each word or phrase.

The mean scale value provides an index of where, on a scale of 1 to 10, each word belongs. For the quality words, for example, it suggests that it is better to be called "outstanding" ($\overline{x}$ = 9.18) than "superior" ($\overline{x}$ = 8.82).

The standard deviation (SD) provides an index of the relative ambiguity of the word. A large SD indicates that people varied more greatly in where they thought the word belonged on the scale.

Differences between words was an important consideration. Whether the difference between the mean scale values of two words is statistically significant can be determined by comparing the upper (UCI) and lower (LCI) .05 confidence intervals (CIs) for the words. This was a part of Long's study. Differences were studied by this technique. If the upper CI for a word is equal to or exceeds the lower CI for the word being compared, we conclude that the two words are not significantly different. In constructing a rating scale, one would probably not want to include two scale anchor points that were not statistically different from each other. For example, in most situations one would want to avoid including "outstanding" (LCI = 9.06) on the same scale with "excellent" (UCI = 9.C7).

Another index of the stability of individual words is the extent to which different groups of people see the words differently. To study this, t tests of differences between various groups were calculated. The general findings were:

1. Professional personnelists, the people who often construct rating scales, seemed to have a common perception of these words among themselves, but it is different from the perception of the people with whom they deal, i.e., clerks, other professionals, and managers.

2. Women interpreted a significant number of words differently than men.

3. Women almost invariably assigned a more positive scale value to words than men.

4. There appeared to be a difference between how the people in public institutions use these words and how they are used in private institutions.

5. Being "average" seemed to be generally perceived to be on the negative side of neutral. Its average scale value was 4.91; on a ten-point scale the "neutral" point would be 5.50.

People who construct or use rating scales for performance evaluation, job analysis, or other purposes can use the results of this study to assist them in preparing meaningful scales. The author has prepared a non-technical guide for using the data to construct meaningful rating scales.

Recent Approaches to Job Analysis

Maureen M. Kaley and Sandra Diaz
Professional Examination Service

The need for compliance with the Uniform Guidelines regarding test validity has created a greater interest in job analysis. Several approaches to job analysis have been used by those involved in personnel selection. A review of approaches was given.

The paper dealt mainly with the role of a delineation approach to conducting a job analysis. First emphasis was on securing task statements that are specific in respect to such questions as: *What* activity did you perform? *To whom* or *to what* was your activity directed? *Why* did you perform that activity? *How* did you accomplish t' activity?

The authors presented various rating scales for evaluating task statements for the purpose of determining their importance and criticality. They discussed how their approach was used to develop and validate test specifications for various occupational groups.

PRE-CONFERENCE WORKSHOPS (Full-Day)

IPMAAC and IPMA provide training conferences and workshops to update personnel professionals about current technical procedures and new approaches to selection, promotion, performance evaluation, and so on. As a part of its efforts in this direction, IPMAAC held four pre-conference workshops in conjunction with the annual meeting. The workshops are summarized below.

CODAP System 80

Leader: Doug T. Goodgame, Texas A&M University

This workshop acquainted personnel professionals who did not have computer background with System 80. System 80 is an extensive series of computer routines designed to process data for job analysis, item banking, and validation. Areas covered included:

- Data collection procedures--the use of task inventories
- A comparison of utility of various job analytic methods
- Basic operating characteristics of CODAP System 80
- Position administration including selection, training, job evaluation and classification.

The participants also were familiarized with the printouts the system produces, their uses in personnel decision making, and the person-equipment requirements necessary to operate the system.

How to Develop and Administer a Structured Oral Interview for Selection Purposes

Leader: Louis M. Laguardia, American Express Company

This workshop was geared toward personnel technicians who have the responsibility for developing and administering structured oral interviews within the context of a selection program. In this workshop, participants learned to develop structured oral interviews for selection purposes, following professional, technical and legal requirements. After a brief introduction to the theoretical framework governing structured oral interviews, participants were taught how the validity and reliability of the procedure are affected by the nature and content of the questions, documentation, structure, controls, and rater errors. The participants had the opportunity to actively engage in a mock oral interview situation.

PRE-CONFERENCE WORKSHOP (Half-Day)

Biographical Data in Personnel Selection: How to Do It!

Leader: Jennifer French, County of San Bernardino, California

Although biographical data have been used successfully as a predictor of work performance and tenure for decades, interest in its use has increased in recent years due to its tendency to produce little or no adverse impact. This workshop focused on the development and validation of a biographical inventory. A brief review of situations in which biodata have been used successfully aided participants in identifying potential applications in their organizations. The resource requirements (e g., sample size and data analysis needs) were discussed, and a basic research design was described ("What data should I collect from whom and what will I do with the data?") Sample response data and analysis examples were used to guide participants through the process of developing a scoring key for a set of biographical inventory questions.

The Application of Content Validation Procedures to Assessment Centers

Leaders: George F. Dreher, University of Kansas
Paul R. Sackett, University of Kansas
Steven D. Norton, Department of Defense,
Centralized Referral Activity

This workshop was designed for persons currently using, developing or considering the development and application of assessment centers and similar procedures. Conceptual literature, federal selection guidelines, court decisions, and empirical research findings were reviewed in exploring whether or not it is appropriate to rely on traditional content validation procedures as the sole justification for demonstrating the job-relatedness of the assessment center process. Alternative validation strategies related to the Uniform Guidelines were presented.