DOCUMENT RESUME

ED 337 476 TM 017 290

TITLE	Proceedings of the 1986 IPMAAC Conference on Public Personnel Assessment (10th, San Francisco, California, June 15-19, 1986).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	Jun 86
NOTE	188p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Assessment Centers (Personnel); *Evaluation Methods; Job Analysis; *Job Performance; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; *Public Sector; Scoring; Test Use
IDENTIFIERS	International Personnel Management Association

ABSTRACT

The International Association of Personnel Management Assessment Council (IPMAAC) is a section of the International Association of Personnel Management devoted to individuals involved in professional level public personnel assessment. Author-generated summaries/outlines of papers presented at the IPMAAC's 1986 conference are provided. The presidential address is "Personnel Assessment: The Next Ten Years" by B. W. Davey. A special presentation is "Where We Have Been and Where We Are Going: An Appraisal of IPMAAC" by C. J. Lindley. The keynote address is "A Valediction for Testing Guidelines" by W. A. Gorham. Twenty-seven papers are summarized under the following paper session titles: "Assessment Center Topics"; "Innovations Related to Work Samples, Simulations, and In-Baskets"; "Attrition: Analysis and Selection-Related Solutions"; "Psychometric Issues and Techniques"; "Unique Public Sector Experiences: Special Problems and Solutions"; "Performance Appraisal: Direct Applications for Selection"; "Microcomputer Administered Testing: Three Approaches"; "Oral Examinations: Unique Approaches to Development, Rating Scales and Rater Training"; and "Selected Papers". Two invited speakers' papers are summarized: "Employee Drug and Alcohol Abuse--Industry's Approach" by P. P. Greaney; and "Touring Performance Appraisal in a Time Capsule" by G. B. Brumback. Outlines of three papers presented during a poster session and two other papers in an untitled paper session are included. A subject index and an author index are provided. (SLD)

IPMA Assessment Council

PROCEEDINGS OF THE

1986 IPMAAC CONFERENCE

ON

PUBLIC PERSONNEL ASSESSMENT

JUNE 15-19, 1986

SAN FRANCISCO, CA

Published and distributed by the International Personnel Management Association (IPMA). Refer any questions to the Director of Assessment Services, IPMA, 1617 Duke Street, Alexandria, Virginia 22314, 703/549-7100.

PROCEEDINGS OF THE TENTH ANNIVERSARY CONFERENCE OF THE 1986 INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL

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

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

The PROCEEDINGS contain mostly summaries and condensations of presentations, but some are more complete than others. The summaries were made by the reviewer(s), and while every attempt has been made to accurately represent each presentation, persons should contact the author(s) directly before quoting results. While many tables and statistical data are included, others had to be excluded because of length. However, bibliographies are included if they were available.

Special thanks go to Jennifer French, San Bernardino County, California, Program Committee Chair and her Committee members for bringing us this professionally stimulating Tenth Anniversary Conference.

PREPARED UNDER THE GENERAL DIRECTION OF:

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

ASSISTED BY:

Thelma Hunt
Professor Emeritus of Psychology
George Washington University

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

Jean M. Shannon, Graduate Student, George Washington University
Charles L. Douglas, Research Associate, IPMA
Mary Ann Diggs, Secretary, IPMA

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;
- assisting with the many legal challenges confronting assessment professionals; and
- coordinating assessment improvement efforts.

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

Published and distributed by the International Personnel Management Association Headquarters:

1617 Duke Street
Alexandria, Virginia 22314
(703) 549-7100

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

TABLE OF CONTENTS

PRESIDENTIAL ADDRESS

Personnel Assessment: The Next Ten Years

Bruce W. Davey, Connecticut State Personnel Department, Hartford, Connecticut

About a month ago, Jennifer French called me and said she needed the title of my Presidential Address right away. So I gave her one, and immediately regretted it. It seemed like a good idea at the time to talk about the next ten years of personnel assessment, with this being IPMAAC's Tenth Anniversary--but after I thought about it a little, I decided that this choice of topic was extremely pretentious. It would be hard to come up with a more pretentious title--unless maybe it was personnel assessment over the next twenty years. I wanted to call Jennifer and change the title, but it was too late because the program was already being printed.

But then I saw Gary Brumback's paper, and I felt much better. If Brumback can cover 4,000 years in his talk I guess I can take a shot at ten years. So here goes.

In considering how predictable the future actually is in this field, one useful exercise is to look at the last ten years, and to ask the question--how much did the personnel assessment field change from 1976 to 1986, and how much of that change could have been predicted? What I think you'll find is a mix--some very predictable trends and some surprises. Also the more general the level of prediction, the more likely it is that the trend in question could have been predicted. For example, some safe bets back in 1976 would have been predictions of increased reliance on data processing methods; increased pressures from various civil rights groups; growing union strength in the public sector; more flexible certification rules; and fewer written tests, but more supporting validity research on those tests. Highly specific predictions within those broad areas would have been more difficult, however. For example, it would have been difficult in 1976 to predict that the comparable worth phenomenon would have taken precisely the form that it did.

When it comes to specifics, a lot has changed in ten years. Let's take a brief look at 1976. In 1976 nobody talked about validity generalization (except perhaps Edwin Ghiselli). Comparable Worth was an unknown term. There were no Uniform Guidelines on Employee Selection Procedures. Differential validity and cultural bias were in vogue. Offices everywhere were devoid of microcomputers or word processors. Who here even knew what a floppy disk was ten years ago? Assessment specialists were still very new to the public sector, and they were almost all working in the area of test validation. The Federal PACE exam was considered to be one of the best-developed and best-validated exams in the country, prior to its demise. And the Intergovernmental Personnel Act was alive and well, and showering us with research funds.

And of course there is IPMAAC itself--brand-new ten years ago and born out of the need for this new breed of public sector assessment specialists to communicate with one another. Things were tough enough then that communication was a matter of pure survival. But if I had to pick out one development of the past 10 years that stands out over all the rest in its significance, I think that would have to be this. (Takes out stickypad) The invention of these little stickypads has totally revolutionized the world of paper pushing and in-basket manipulation. And I don't know if we could have predicted this innovation.

That brings us from 1976 to 1986. What about the future? Obviously, that's a little tricker, and for the most part, for reasons discussed earlier, I'll have to be pretty general to be effective in my forecasting. However, I would like to make a few specific predictions.

In 1996, I predict that IPMAAC's President will be someone named Deborah von Pallenburg. Deb at that time will be the Chief Personnel Psychologist for the Federal Office of Big Government. She'll win handily because she'll have the support of all those federal government personnelists who came into the system with the return of big government in 1992...and a few people will express concern that the federal members are beginning to take over IPMAAC. Remember that you heard it here.

I also predict that in 1996 the IPMAAC Hacker will celebrate its 1,000th page anniversary with a software offering that creates a three-dimensional moving hologram of Larry Jacobson drinking pop and Bruce Davey drinking beer and both blowing out candles on an anniversary cake. That's sort of an in-joke for IPMAAC Hacker fans.

Some other predictions of a less specific nature will now follow. Some of these will relate to the direction in which I think the roles of the assessment specialist will evolve; some will relate to specific assessment trends; and some will relate to technological advances. I have a feeling my predictive accuracy rate will be best in the area of technological advances.

I have to begin with my favorite area of prediction--trends in computerization. It is a safe bet that the computer will become even more integral to our work than ever before. The computer, and especially the personal computer, has arguably brought about some of the most significant changes in personnel assessment work between 1976 and 1986, and it seems certain that this trend will continue and will accelerate. I believe that we are on the threshold of yet another computer explosion.

There's a new computer technology on the horizon that's going to keep this incredible revolution in an acceleration mode for a long time to come. The computers of the past (and doesn't that sound strange, to be talking about the computers of the past?)--and the computers of today can process bits of information at literally lightning-fast speeds--but they're limited by the fact that they presently can only process one instruction at a time. That creates a bottleneck known as the von Newmann bottleneck. Computers may have awesome processing speed which far exceeds the calculating capabilities

of the human brain, but they have been unable to compete with the human's brain's ability to simultaneously process a lot of information, in parallel. We can process sights, sounds, sensations, and thoughts all at the same time. A computer can't do that. It has to stack its bits of information one behind the other because of the von Newmann bottleneck. BUT--a new computer technology is developing right now which permits truly simultaneous processing of an incredible amount of information. The first parallel processing machines are now being tested and are passing with flying colors. So, the computer revolution accelerates onward, and once again, it's going to be a new ball game. Research on artificial intelligence is going to blossom with parallel processing, and the onset of true "thinking" computers is going to become a reality in the next ten years. That's not a prediction, it's the recognition of an inevitability.

With computers that powerful, the nature of the interaction between computers and humans is going to change. Computers are going to become effective at recognizing speech and sounds and visual information, and at speaking themselves. I'll let your own imagination consider the possibilities of that, both for the world of work and information processing in general, and testing in particular.

Now I'm going to backslide to the more conventional type of computer and its role in the immediate future of personnel assessment. I see it becoming more tightly integrated to a number of aspects of personnel work, more so than ever before. For example, in testing, the microcomputer is likely to be used more and more for test administration. The micro-computer is capable of setting up a much more personal interaction with test-takers because it can give each one individual attention, and it can supply quick feedback. For those of you who went to the symposium on approaches to microcomputer-administered tests, you know that there are a variety of new testing techniques possible with a microcomputer which are not possible in conventional modes. Computerized adaptive testing allows the computer to identify a candidate's ability level with about seven times more efficiency and speed than a traditional test. Simulation testing allows subjects to make decisions, and then gives them feedback on the consequences of their actions, and lets them continue to work through the problem in their own way. Candidates are more accepting of these kinds of tests than they are of traditional multiple choice tests, because they can see the correspondence to reality, as opposed to the answering of a bunch of multiple choice questions and getting feedback on their performance a month later. Candidates want feedback, and computers can provide it.

I see the computer being more effective in other areas of personnel as well. One possible fruitful area is that of performance evaluation. Perhaps, somehow or other, the computer can be the focal point of a more effective performance evaluation system. I can visualize a setup in which there's an interaction between the computer and the evaluator, with the computer giving feedback on the rater's tendencies, or inconsistencies, or how the ratings on the employee being rated compare with all others throughout the department, and so forth. The computer might aid in fashioning a better narrative description of performance as well. The interaction might even be such that the rater supplies the narrative and the computer converts

that into a numerical rating (if one is necessary). In other words, I can see a situation where the computer could ask the supervisor questions about the worker's performance, or give the supervisor a lot of choices from which to select appropriate responses, which the computer would then convert into a series of ratings. Somehow I think there would be fewer errors in performance ratings if the supervisor completed the performance rating exercise under the counseling of someone else--even if that someone else is a computer.

I could talk about computer applications all day--but that's enough for now. I'd like to talk now about what I see to be the changing role of the personnel assessment specialist over the next ten years.

Personnel Assessment Specialists in the public sector are a fairly new breed. They were very rare in the public sector in the 1950's and 1960's. They seem to have arrived as a common fixture in the early 1970's--not coincidentally, at about the time of the EEO Act of 1972, which extended the jurisdiction of EEOC's Testing Guidelines to state and local governments. At that time, personnel assessment specialists were primarily engaged in testing and test validation, because that is where the greatest perceived need was.

That trend seems to be changing. Personnel Assessment Specialists are working their way into other areas of personnel where they are needed--for example, classification and compensation. This is in part due to the traditional linkage of Assessment Specialists to the job analysis process, and in part due to the pressure that the comparable worth movement is putting on the compensation function. The comparable worth movement is placing the same kinds of pressures on the classification/compensation staffs as was placed on test development staffs over the previous ten years. And the skills required to meet the challenge are again those of the assessment specialist, especially now that they have court experience. It also seems clear that the talents of assessment specialists can be put to good use to design more sophisticated and scientific approaches to salary surveys than are now typically done.

In fact, there are many places where the assessment specialist's skills can be used and should be used: Attitude surveys; Training needs analysis; Productivity measurement; Analysis of sick leave and turnover data; Development and implementation of more sophisticated performance evaluation systems; and, perhaps, pay-for-performance systems.

What I see happening in the public sector is that personnel assessment specialists are serving strictly as testing specialists to a lesser and lesser degree and becoming personnel assessment generalists to a greater and greater degree.

A comparative look at IPMAAC Conference agendas over the ten years of its existence will clearly confirm the trend. In IPMAAC's early days, IPMAAC's program was almost entirely testing. Now, it is a cornucopia of assessment practices.

It appears that in the public sector, personnel assessment specialists are becoming much more like the classic conception of I/O Psychologists. What I find especially interesting about this is that we're on our way to coming full circle on the specialist/generalist continuum. As assessment specialists are becoming more generalized, personnel analysts have become more specialized. We and they seem to have passed going in opposite directions.

Maybe that needs clarification. Ten or fifteen years ago it seemed that every centralized personnel department operated on a personnel generalist concept. Over time, that has shifted, especially at the state level. More and more states have gone to a specialized approach, with the examination section being split off from the classification section. In fact, many states have specialty units within their testing operation.

So, while personnel analysts have gotten more specialized, we personnel assessment specialists have taken on broader and more varied responsibilities. Maybe we should start calling ourselves personnel assessment generalists.

So much of what we do, and what our employers want us to do, is shaped by outside forces--forces like EEO, and comparable worth, and truth in testing--that a discussion of the next ten years would be barren without speculation on what sorts of forces will be pressuring us in the future.

Comparable worth is a major force which is just starting to hit its peak, the comments of government personalities notwithstanding. I think the issue of female equality in the workplace will continue to grow as an issue on the late 1980's and early 1990's--it is not going to go away just because some members of the present administration want it to. It's too big an issue to go away.

Another group which is more likely to exert its rights as time goes by is the candidate group at large. They have started to do that on college entrance exams, and I can't think of any reason why they wouldn't extend that, in time, to employment tests. All the indicators are positive. There's the Truth-in-Testing movement which hit heavily in the college arena; the Freedom of Information movement; and a general trend toward consumer advocacy in America.

In addition, labor unions in the public sector continue to establish themselves, and one of their traditional issues is exam disclosure. All the signs seem to point toward more complete disclosure of test information. It's a challenge that the personnel testing field will have to respond to. We can't sit back and wait for the issue to engulf us.

If we go into an economic boom, test disclosure won't be as much of an issue. Maybe the way we need to respond to pressures for full disclosure on tests is to provide more feedback to candidates, before and after the test. We can tell them what to expect and how to prepare for it. And afterward, we can give them more feedback on why they got the score that they did and what that score actually means. On a recent consulting project where candidates were looking for full disclosure of the test so

they could learn from their mistakes, as they put it, I instead gave each candidate a breakdown of how well they did on each exam subtest, and also how the candidate group at large scored on the average on each subtest. They were very happy with that. And I think we're going to have to systematically do more of that in the future if we're going to successfully deal with pressures for disclosure.

An important trend to consider is that the work force is getting older on the average. The baby boomers are aging.

Why did I say that? Now I'm depressed.

I think that the trend towards an older work force is going to lead to an aggressive push for the rights of older workers, and I think the chief points of attack will be selection, performance evaluation, and promotion. And there is potential for the same kinds of knotty psychometric and social issues and the minority adverse impact issue has produced.

Think a minute about what happens to a worker, regardless of age, who isn't very good. He or she stays at a particular job level, and gets older. If there's low turnover, other than promotion, after a while you'll have two sets of workers in that job...people who have been around a while and not promoted because they weren't very good and never <u>were</u> but now are older and not very good...and young turks. The young turks get promoted and this leads to adverse impact.

You'll notice that I've been talking for perhaps fifteen minutes and haven't yet mentioned validity generalization. Now why is that? I guess it's because I seem to have an approach/avoidance reaction to validity generalization. For a long time I wasn't sure why, but it's finally clear to me. I think we in the testing field owe a lot to Frank Schmidt and to John Hunter, and to validity generalization and utility analysis, because they came along at just the right time. Testing was under fire, and these guys and some others came along and said, "Hey--wait a minute. We've got data to show that tests work, and that basic ability tests are valid across a wide spectrum of jobs, and using them can save you money." The testing field needed to hear that, to give it back some confidence at a time when it was being attacked from all sides.

In that vein, validity generalization was great. But on the <u>avoidance</u> side of my approach/avoidance complex, I'm concerned that this movement might inadvertently have within it a call for testers to "stand pat." Let me be clear that I'm not saying that this is the position of Schmidt and Hunter... but many practitioners seem to believe that the basic ability test is the be-all and end-all of personnel selection, and you can't improve on it, and instead we need to stand behind all the research that has been done on its effectiveness. It's as if VG proponents are saying "Don't worry--what you've been doing is fine." Well, that has a very conservative philosophy if you think about it, and I'm not very conservative. It has within it the seeds of a stand-pat position, and that's my big concern. We can't stand-pat at a time when the best of the traditional tests predicts perhaps 25 percent of the variance in job performance, and is unpopular as hell besides.

Let's reflect for a moment on the unpopularity of the written ability test. At the same time federal government was doing some of its validity generalization work on the PACE exam, it discontinued its use. There's a message in there somewhere, and I think the message is that validity is extremely important, but adverse impact and candidate acceptance need to be considered too. Otherwise, you lose.

I'm hoping that future tests will look at people more multidimensionally. There's a lot about human potential that we don't yet understand. I hope we're going to get a lot better at measuring it, and at cutting into that 75% of the variance that we can't predict. Again I think the computer holds part of the key to doing that.

And how could I possibly sit down without talking about what the next ten years holds for IPMAAC? Well, I already told you about Deborah von Pallenburg and the thousandth page of the IPMAAC Hacker...now I'll give you my more general predictions.

One feeling I have, and which I alluded to earlier, is that IPMAAC's membership composition will get more similar to Division 14 of the American Psychological Association. Each year we seem to increase the percentage of consultants and university-based members. I've already heard some people refer to IPMAAC as a "poor man's Division 14." I prefer to see Division 14 as a "rich man's IPMAAC."

There are things that I hope will make us remain unique as an organization. Chief among these is the cooperation spirit of IPMAAC. I see that as one of IPMAAC's defining characteristics and I hope that spirit will never fade. There are lots of people in this organization who feel that the way to advance our profession is through shared products and shared technology and shared communication and support. And they're right. I hope that as IPMAAC continues to mature as an organization, it never loses sight of this fundamental concept--because it is the foundation and spirit of this organization.

It's been a great honor to serve you as your president for the past year. Thank you for the opportunity--Thank you for your support--and make you reservations early for Philadelphia.

* * *

SPECIAL PRESENTATION

Where We Have Been and Where We are Going: An Appraisal of IPMAAC

Clyde J. Lindley, IPMAAC Historian, Center for Psychological Service, Washington, D.C.

"Coming together is a beginning,
Keeping together is progress,
And working together is success."

Theodore Roosevelt

INTRODUCTION

This is a very apt quotation for beginning this paper. IPMAAC began by a coming together of varied persons working in the personnel field with special interests in the area of assessment. These persons were varied in educational background and the nature of their work experiences. They were held together by their common interest in problems of assessment of persons in the workplace. They have kept together with increasing strength and numbers over IPMAAC's ten-year history. The diversity of their backgrounds has given more challenge to their approaches to projects undertaken. And if working together is the measure of success it has been attained in large measure.

The title of a talk or paper is always an interesting consideration. Sometimes it is invented after the paper is written to fit the words set down. Sometimes it is there as the starting point. The latter is the case for me. I selected the topic and I'm stuck with it.

As I examine the topic, "Where We Have Been and Where We Are Going: An Appraisal of IPMAAC," I am first impressed by its indication that IPMAAC is a going concern. How could we have been someplace without being? How could we be going someplace without surviving?

The "appraisal" part of the topic suggests that we are mature enough to take a critical look at what we have been doing. This critical look should assess those accomplishments of real significance to the purposes of IPMAAC as well as the identification of our shortcomings or areas of needing improvement. This appraisal should end in helping us formulate better defined goals with some indication of their importance and priority.

OBTAINING THE HISTORY

Let us now look at the process of obtaining the history. In May 1982, I submitted a report for IPMAAC's Long Range Planning Committee titled "Looking Backward in Order to Look Forward." At that time my objective was

to review what had been accomplished by IPMAAC with an assessment of how well past plans have been carried out so that we can perhaps better target our future goals. This task embodied the review of all Minutes of IPMAAC Board Meetings, Committee Reports, Newsletters (ACN) and related documents, and included discussions with key personnel. As your Historian, I have continued a similar process, with perhaps just a little more attention toward the founding of IPMAAC. This involved a review of IPMA's Executive Council Minutes also, and discussions with the Executive Director of IPMA.

Before I summarize the early events that led to IPMAAC, let me comment very briefly on this process of evaluating and documenting our history. I found this activity to be highly stimulating. Also, I continue to be impressed with the extent and breadth of Committee activities and the high professional standards of all those who have been directing and guiding IPMAAC and selecting targets for accomplishment. So many persons have contributed their time and efforts to this process that it would be impossible to mention them all. So as I present this historical perspective, please realize that there are many unidentified contributors in the background. Many persons who served on the Board of Directors throughout our ten-year history, and the persons on IPMAAC's Committees are our unsung heroes. Their dedication to IPMAAC will be obvious when I cite our accomplishments.

ORIGINS AND ESTABLISHMENT OF IPMAAC

Early in 1975 IPMA began planning more concrete ways to meet the needs of members with special interests in selection and in other areas. This culminated in conducting a Symposium for Selection Specialists in Chicago at the Water Tower Hotel, July 6-9, 1976. About 154 persons attended this meeting. Thomas Tyler was Director of Test Services for IPMA in 1975, and along with Donald Tichenor, Executive Director of IPMA, started the ball rolling by inviting comments from William Gorham, Director, Personnel Research and Development Center, USCSC, and Charles Sproule, Chief, Division of Research and Special Projects, State of Pennsylvania. Here I would like to emphasize the significant role played by Tom Tyler. Through his efforts he encouraged and stimulated the development of the selection symposium and provided an opportunity for the persons in attendance to consider how they wanted to meet their unmet needs. Bill Gorham, Charlie Sproule, Ted Darany, and Glenn McClung, to mention only a few, had key roles in this development.

The ideas about the new organization were discussed at the Chicago meeting where forty persons worked on special committees related to this organization's development. An ad hoc executive committee was formed to establish the new organization within IPMA. It was to be called "IPMA Assessment Council." The temporary executive committee's function was to guide the work on developing the new association, make plans for membership, annual meetings, etc. The committee was chaired by Bill Gorham. Members of the committee were: Andy Anderson, S.C. Personnel Division; Theodore Darany, (at that time) USCSC, Warminister, PA; Charles W. Grapetine, Milwaukee Personnel Department; James C. Johnson, Tennessee State Department of Personnel; Arleen Kleber, CODESP, Garden Grove, CA; Glenn McClung, Denver Career Service; Robert Shoop, MO Personnel Division, and Charles Sproule, PA

State CSC. Again let me repeat that many other persons contributed to the activities of this Committee. Some are mentioned in the Special Assessment Council 1986 San Francisco Conference issue.

The Executive Council of IPMA at its meeting of October 18-19, 1976 requested President Muriel Morse (IPMA) to respond to Dr. Gorham's request advising that IPMAAC was approved as a section of IPMA.

We were born!

THE ACRONYM "IPMAAC"

Let me digress briefly on the acronym IPMAAC. The last two letters "AC" stand for Assessment Council. These distinguish us as an organization. The Assessment Council is a section of the International Personnel Management Association (IPMA), the parent organization. Those making up the original council represented a group of IPMA members particularly interested in psychological testing and its application to such personnel problems as selection, placement, promotion, performance evaluation, etc. The term "assessment" was chosen rather than "testing" to better indicate the broader coverage in terms of functions and types of assessing procedures used. In addition to psychological tests, we shall keep in mind that assessment procedures also include such things as interviews, training and experience rating, self-esteem, physical examinations, strength and agility testing, assessment center evaluations, and evaluations of the functioning of the public service organization itself, usually referred to as organizational development and management. Throughout all of these assessment concerns runs the concept of ethical standards for practice and professional accountability. This broadened coverage of assessment - grown out of testing - has markedly increased the responsibilities and importance of IPMAAC.

IPMAAC PURPOSES

It may be helpful at this point to review the purposes of the International Personnel Management Association Assessment Council. (This is taken from the November 1977 IPMAAC Assessment News.)

1. To support the general purposes of the International Personnel Management Association.

2. To encourage and give direction to public personnel assessment maintenance and improvement efforts in fields such as, but not limited to, selection, performance evaluation, training evaluation, and organizational effectiveness.

3. To encourage and facilitate intergovernmental cooperation, information exchange, and resource sharing.

4. To define professional standards for public personnel assessment.

5. To encourage, give direction to, and provide means for the delivery of training and education efforts to upgrade the expertise of public personnel assessment specialists.

6. To influence public policy relating to public personnel assessment.

7. To heighten the awareness of public officials and administrators of the needs of public personnel assessment.

IDENTIFICATION OF PROBLEM AREAS

Over the years there have been many areas that IPMAAC Boards and/or Committees have emphasized again and again that are in need of greater progress or which represent deficiencies. I have grouped these in three areas.

Membership in IPMAAC

1. Continuing need to attract new members.
2. The need to attract minorities.
3. The need to reach smaller public service agencies that have few resources in the assessment area.

Communication links

1. Too little communication to the membership.
2. Not enough communication with personnel directors.
3. Too little communication among committees.
4. The need to provide continuing communication with IPMA.
5. The need to promote information sharing.

Professional identity

1. The unique role that assessment specialists have in public personnel assessment.

2. The broad and varied backgrounds of persons in IPMAAC.

3. The emphasis on practical but professionally sound approaches to solving assessment problems.

4. The problem of developing professional standards for the wide variety of persons engaged in assessment activities in public personnel work.

MAJOR ACCOMPLISHMENTS

Now I would like to talk about our major accomplishments. It would be impossible to cite all the accomplishments. However, here are what I consider to be major accomplishments achieved by IPMAAC since its founding in the Fall of 1976:

- A viable IPMA Assessment Council with about 500 members. You can be proud of your organization for it started out by providing professional information exchanges in the assessment area and continues this direction now. We are truly a professional organization!

- Sponsor of Annual IPMAAC Conferences.

- Publication of IPMAAC Newsletter (IPMA Assessment Council News); initially three times a year, now quarterly, with expanded coverage, and regional correspondents.

- Sponsor of Workshops and Seminars (at the IPMAAC Conference, at Regional IPMA meetings and at the IPMA Annual Conference); sponsor of program sessions at the Annual IPMA Conference.

- Development of Standards for Sharing Item Bank Materials.

- Publication of the Proceedings of Annual IPMAAC Conferences on Public Personnel Assessment.

- Publication of Sourcebook: Information Sources and Services in Personnel Assessment (two separate editions, 1981 and 1983).

- Sponsorship of a Student Award Program (first one at the Annual IPMAAC Conference, June 6-10, 1982, Minneapolis, Minnesota).

- Completion of a survey of public sector agencies nationwide to identify common research needs, successful cooperative projects, and useful sources of information on personnel assessment.

- Publication of the IPMAAC Hacker as a special resource to persons actively using computers in some phase of personnel work.

- Publication of Personnel Assessment Sources (PASS) on a regular basis.

- Code of professional principles (ethics) for personnel selection specialists.

- Review of the Uniform Guidelines on Employee Selection Procedures.

- Review of the revised APA Standards for Educational and Psychological Testing.

- Nationwide job analysis of selection specialists (ongoing, with substantial progress already made).

RECOMMENDATIONS: FUTURE OUTLOOK

These recommendations point to things that in my opinion should occupy thinking and efforts on the part of IPMAAC. They are particularly addressed to all IPMAAC members. They are not to be looked upon as offered in a spirit of negative criticism or of neglect in appreciating the many excellent accomplishments and services of IPMAAC, but as possibly helpful suggestions for charting future emphases.

1. IPMAAC should continue to strive for a more effective relationship with IPMA.

   IPMAAC was founded on an organizational structure in which IPMA constitutes the parent group, and IPMAAC constitutes a subgroup organizing itself in relation to the parent group. IPMA strongly supported the subgroup's organization and purposes, and the beginning relationship was obviously strong. Over its ten-year history, the relationship at times seems to have grown more tenuous. In later years relation have improved. It was good to hear Dr. Pounian (President of IPMA) state in his opening address, "IPMAAC is an essential part of IPMA." He emphasized that there is "a need to develop and strengthen that relationship." Although IPMAAC represents specialized interests, it has much to gain by being a part of the larger area of personnel interests represented by IPMA. Hence, the recommendation that as the Assessment Council continues to grow, it consider maintaining effective communication with IPMA as being very important. Here are a few suggestions for IPMAAC to consider:

   a. Invite selected IPMA members to make presentations at IPMAAC Annual Conferences. (They should not be IPMAAC members).

   b. Involve personnel directors in discussions about possible joint projects.

   c. Strengthen interchanges by inviting more members of IPMA Executive Council to be present at meetings of the IPMAAC Board of Directors.

2. Organizational Functioning Needs Constant Reviewing.

   In my earlier report I emphasized communication in organizational functioning and I reemphasize it again now.

   Several important Long Range Planning Committees or Continuity Committees have intensive analyses of IPMAAC's objectives and recommended specific actions to improve our organization. Some of these have focused on IPMAAC's Board functioning, improving IPMAAC's financial management, and its professional recognition and stature. Much of this activity has resulted in real improvement in direction and identifying practical goals for the organization. This type of activity should be continued with the opportunity for more input from the membership to consider the various objectives and/or goals.

Organizational functioning, including short and long-term planning, are highly dependent upon communication. So it follows that long range planning for IPMAAC must give attention to the communication problems. We must be sure that IPMAAC membership is informed of what is going on and is sufficiently brought into the picture. We must be sure that our organizational structure of committees for carrying out our functions is effective. Too many committees and committees too large to meet face to face are likely to bog down because of communication problems.

Before a major activity (policy) is implemented one should always ask the question, "How will this impact upon the major goals of IPMAAC? How will this affect our efforts at recruitment and retention of members? How does this action impinge on IPMA?"

To maintain progress, the long range planning effort must, as it were, have its eye in the sky, must be more attuned to satellites than the metallic telephone wire. IPMAAC must become a prophet in sensing needs and services in the personnel area of the future. Target objectives here must first be exploratory. They must be brainstormed with all stops off for the expression of ideas. The long range problem is really one of defining and clarifying objectives for the future. Even an organization like IPMAAC usually spends most of its resources on "putting out fires," rather than effective long range planning. Therefore, it is important that some concerted attention be given to developing what the future objectives of IPMAAC should be, as the world of work changes about us.

3. IPMAAC NEWSLETTER (ACN) continues to need support.

The production of the Newsletter is an important function of IPMAAC because its purpose is to let the membership know what is going on. At present the Newsletter does not convey enough information on the activities of the IPMAAC Board of Directors and its various committees. Going back to the IPMAAC Board of Directors' Meeting in April 1977, these comments were made about the Newsletter. It should provide the official organ for transmitting IPMAAC business and correspondence to IPMAAC members. It should serve as the vehicle for announcement of IPMAAC activities and other activities of interest to members. I should also promote membership and provide news of the activity of IPMAAC members. This was a good statement of purpose. To summarize, the ACN should cover the significant activities of IPMAAC, its Board of Directors and committee chairs/members. The Regional Correspondents need your help in submitting information about your activities in their regions.

4. Membership and membership involvement must be recognized as very important.

IPMAAC membership is now somewhat below its peak. (We should be concerned with what this means.)

There are some membership areas which are sadly lacking. One of these is colleges and universities. There is a need for recruitment of members among academic personnel who teach courses in personnel and furnish an important aspect of training for potential and actual personnelists.

IPMAAC sorely needs a separate accurate list of members. This membership list should contain at minimum the name, agency or organization, position title, address and telephone number. A first-time separate publication is recommended which thereafter could be published in the IPMA Membership Directory. There should be a separate listing of IPMAAC members in the IPMA Directory.

Adequate membership from minority groups must be an aim. Special attention needs to be given to minority groups (blacks, Hispanics, Asians), dependent to some extent upon the locale of functioning.

5. IPMAAC Annual Conferences.

The annual conferences should be looked upon as a most important activity contributing to the survival of IPMAAC. The excellent quality of the conferences so far has undoubtedly had favorable influences in creating a good image for IPMAAC.

Here it is desirable to repeat prior recommendations (my original report of the Long Range Planning Committee in 1982).

Conferences should offer appeals to both technically trained and non-technically trained personnelists. Programs should be varied, especially in the direction of presentations of benefit to persons new to the assessment field. Conferences should be planned to attract non-members in the personnel field as well as members. Attracting them might constitute a road to their becoming members.

Efforts should be continued to encourage members to attend the annual conferences. IPMA records indicate less than one-half of membership attends. It would be helpful, in planning efforts to improve attendance, to study in more detail the reasons for non-attendance. With increasing cutting of agency support of employee conference attendance expenses, these factors related to non-attendance need reconsideration and serious attention. Perhaps there should be consideration of a request from members to support contributions that would be used to send a limited number of younger new members to the annual conference.

6. Contributions to "Public Personnel Management."

The recommendations made in the past that IPMAAC should become more visible in IPMA's journal should continue to be emphasized. The past recommendations were brought to fruition in the Special IPMAAC Issue of the journal (Winter 1984), published in 1985. This issue was

devoted to Assessment Techniques and Challenges, with myself and Thelma Hunt serving as special issue editors. IPMAAC might consider recommending to IPMA another special issue on an appropriate topic.

I bring up again the occasionally recurring question of IPMAAC undertaking publication of a separate journal. This would be a very expensive undertaking, and does not seem to be currently justified as best serving IPMAAC's needs. Such a journal would also be in competition with already well established journals dealing with assessment and measurement issues (see the Sourcebook: Information Sources and Services in Personnel Assessment).

7. Relationship with Other Organizations.

IPMAAC should continue relationships with professional organizations functioning or contributing to the assessment field. In the past few years ties have continued with PTC, WRIPAC and other consortia, and we have strengthened relationships with Division 14 of the American Psychological Association and these activities should continue. Liaison with APA's Division 5, Measurement and Evaluation is also important. But IPMAAC should not become so identified with such Divisions of APA that their objectives are indistinguishable. All IPMAAC members are not professional psychologists. IPMAAC must continue to maintain a broader spectrum of membership.

Some areas of IPMAAC concern call particularly for closer relationship with other organizations. As an example, I think of the area of career development. Here the American Association of Counseling and Development might be interested in liaison activities.

AREAS NEEDING EMPHASIS

Some important areas of broad personnel concerns appear neglected by IPMAAC. I will mention only a few.

1. The importance of motivation in relation to employment.

To apply for a job one must be motivated by knowledge of its nature and opportunities to satisfy one's potential. To stay in the job one must be motivated by some "reward" (ranging from money alone and something to occupy one's time, to highest level of self-actualization). Public employment has been accused of being the place where un-work-productive motivation and lack of self-actualization have been able to flourish. In anticipated "tight" public money and increased legal restraints on retentions and promotions, motivational aspects of employee qualifications are likely to become much more crucial in hiring and promotion. IPMAAC can make real contributions in this area.

2. The problem related to retirement.

IPMAAC's potential contributions cannot be set down in detail. This would have to be place by place and agency by agency. From my observations of retirement approaches and systems, one general recommendation comes out first. Pay more attention to the process of retirement, as contrasted with the "clerical" details connected with effecting it. By the "process of retirement" I refer to informing and preparing the retiree, dealing with the attitudes of workers at all levels toward retirement policies, helping retirees adjust to retirement, etc.

3. The special problems of the older worker.

This problem has been addressed in your Newsletter, the ACN. Many questions remain unanswered, and little data is available in the public employment area (except for the Federal Government). How long can older workers remain productive? How will agencies provide "upgrading" incentives for younger workers if older persons remain in key positions? A real challenge exists in this area.

4. The special problems of women in the workforce.

These range from the long-standing ones related to hiring, promotional, and pay differentials with respect to sex, to newer ones tied in with individual sexual behavior and practices and sexual harassment in the workplace. The propensity, in the present era of legally oriented attitudes, to pursue such issues with legal challenges or lawsuits has emphasized many of the problems that still exist.

The older hiring, promotion, and pay differential problems mainly center around fairness. Does it represent fair and equal opportunity consideration that only a small percentage of police jobs are filled by women? Is it fair that routine office jobs (often considered boring) are mainly filled by women? Is it fair that top management jobs are mostly filled by men? There are many subsidiary problems (to be solved first) before solution of such problems as these can be logically attacked. The most fundamental is the establishment of job tasks and qualifications for performing them. These must then be related to inherent differences between the sexes. If women inherently do not possess a needed qualification for a specific job (such as great upper arm shoulder strength) then differentiation in sexual hiring rates is defensible. Top management jobs have often been discussed in relation to women. Related factors to be considered are the matter of opportunities or lack thereof for women in attaining necessary experiences for top jobs.

5. There must be an awareness of the future changes in the composition of the workforce and the implications for assessment activities.

We need information about the changes in the workplace brought about by the advances of physical and social sciences. The former have

brought about the workplace changes associated with the computer and all its accompaniments. The latter have replaced rigidity with flextime work hours, and quality circles and participatory management emphases. Similar changes will accelerate in the next decade. We should be in the forefront of developing the best methods of adapting to these changes to achieve continued productivity in the workplace.

6. IPMAAC should strive to improve the acceptance and image of public employment.

Public employment needs to be a top goal instead of a last resort. Efforts toward improvement can be directed toward both personal attitudes and the public work environment itself.

FINAL APPRAISAL

In reviewing and evaluating IPMAAC in its ten-year history, it is obvious that the Assessment Council has achieved professional status and recognition in helping solve important problems related to assessment in the personnel field.

IPMAAC has been blessed with good direction by a large number of dedicated Board and Committee members. They have charted objectives and directives for obtaining goals which have produced good results. There is no reason to recommend that IPMAAC adopt any major about face policies.

In the main, my comments relating to evaluations and recommendations have already been given some attention by IPMAAC. Even though IPMAAC has been going in the right direction, it is desirable periodically for any organization to take a good hard look at what it has been doing in order to assess where it might make improvements. My evaluations and recommendations are offered to meet this need.

Working together we can continue our progress.

* * *

KEYNOTE ADDRESS

A Valediction for Testing Guidelines

William A. Gorham, Ft. Lauderdale, Florida

(First President of IPMAAC, 1976)

Almost ten years ago, in Chicago, on July 6, 1976, I addressed the Selection Specialists' Symposium Conference. That, as it turned out, was also the organizing conference for what emerged as this organization: The International Personnel Management Association Assessment Council.

That was the first time that I had been honored to be a "Keynote Speaker." In order to know what was expected of me, I had looked to the dictionary to find out what a proper "keynote address" was, or what a "keynote speaker" was supposed to do. I reported that definition to you then, but in case some of you have forgotten, or weren't there, or wonder what I'm supposed to do today, here it is again:

> "Keynote address or keynote speech. n: an address (as at a political convention) intended to present those issues of primary interest to the assembly but often concentrated upon arousing unity and enthusiasm. [The keynote address...is a highly emotional performance-D.D. McKean]"

As I reread that 1976 address in preparation for this, my second "keynote address," I searched for evidences that I had lived up to, or in this case spoken up to, the definition. Some of the major issues that I presented in 1976 (I'm not sure whether they were those of the conference attendees or my own) were:

- The status of the issue of different group mean test scores and its meaning for us.
- Adverse impact vs. validity. Could validity be expected to overcome fatal cases of adverse impact?
- Evaluating the worth of selection tools to and for our own employers.
- The need for a new organization to meet the emerging requirements of public personnel measurement specialists.

In my own retrospective view (also known as "hindsight," a well-known psychological construct) some of the key issues of that day seem to have been identified. Conference participants, however, added many more in the symposia and papers.

As to whether unity and enthusiasm were aroused, I can hardly claim to have "concentrated" upon that aspect of keynotership, since those of you who were there listened to 26 pages of text before I came to this:

> "We are...proposing a new organization to accommodate the needs of all of those who want to identify as public personnel psychologists... a constructive response to the crises of our time. The time is right; the need is here; we have an opportunity to fill a gap and to provide leadership in our own field. Many of us are enthusiastic and ready to unite. Let us begin to meet our crises together."

It is one thing to arouse unity and enthusiasm, but that is a barren exercise if results do not occur. If political parties don't elect officials then generating unity and enthusiasm in keynote addresses may be fun, but has little other validity.

But today's IPMAAC clearly has continued the unity and enthusiasm for these 10 years. Further, there are results in the form of unique professionally responsive contributions by members and the organization to the common good. Quite simply, you have succeeded. Professional gaps have been continually and ably filled. A new leadership emerged and is well established. I applaud and congratulate you.

Adlai Stevenson in addressing a group once said, "I understand that I am here to speak to you and that you are here to listen. I hope that we both conclude at the same time." If what I have so far said sounds like the end of a keynote address, it is not. Please do not conclude your listening. I still have the obligation and intent to speak about some of today's issues, and I shall, although clearly out of practice, attempt to arouse unity, although I am skeptical that it is needed. Bear in mind that I have not been in the crucible of national issues since 1979. However, seven years may have allowed me to acquire a certain amount of detached perspective.

Besides the founding of IPMAAC, what was happening in our field ten years ago?

- Washington v. Davis was decided by the Supreme Court as we convened. We discussed it at a general session. You will recall that the case involved testing practices in the District of Columbia Police Department. As it turned out, the acceptance of training success as a criterion was probably the most important outcome since Federal enforcement agencies had continually rejected that idea in work on testing guidelines.

- The Federal Executive Agency Guidelines were published two months after we met, in September 1976.

- EEOC withdrew from the Guidelines consensus process, republishing their 1970 Guidelines.

How I managed to talk to you in 1976 about issues of the day without a single reference to impending Federal Testing Guidelines is a total mystery to me today. I had been deeply embroiled in that activity for years. Perhaps so long that I didn't believe that we'd ever conclude. For, a process which should have proceeded along systematic cooperative lines among Federal government agencies was, instead, more nearly like what we envision arms control negotiations to be like. There was more acrimony than harmony; more divisiveness than cooperation; more dependence upon intuition than upon science. I am distressed even today that skilled human resources—including my own—which could have been doing more about the basic problems in minority unemployment were, instead, sapped over a five year period to produce a document in 1978 which has had little influence on the employment of minorities and women.

This is not going to be a "kiss and tell" history session, but a sort of history lesson which I urge you to attend. We may cycle around again someday. When I was in graduate school I was least interested in the history of psychology. I now understand why it is so important. I did not suspect then that I would chronicle and be a part of it; but we should certainly not repeat our mistakes.

What was the problem? Rather, what were the problems? Beyond the sharply contrasting viewpoints of the Federal agencies involved, there was, from the beginning, a lack of acceptance of a sound scientific basis for the development of the technical aspects of the Guidelines.

Now, I must go back even further. Two decades ago, in 1966, EEOC published its first guidelines consisting of some very general principles and a four page report from a three-person "panel of outstanding psychologists, all of whom have broad experience in the testing field..." and an attorney. Now, the usual training and experience which is relied upon in qualifying outstanding psychologists is that of industrial or measurement psychology. One psychologist was a Fellow of Division 14. Excellent! The second was a diplomate in clinical psychology. The third was apparently not a member of the American Psychological Association. Thus, one of four qualified scientifically.

Among other things, the 1966 guidelines stated:

> "g) Tests should be validated for minorities. The sample population (norms) used in validating the tests should include representative members of the minority groups to which the tests will be applied. Only a test which has been validated for minorities can be assumed to be free of inadvertent bias."

I can think of no better word than bugaboo to describe the above requirement. Webster defines that term as

> "1. An imaginary hobgoblin or terror described to frighten children into good conduct.
>
> 2. Something that causes needless fear."

That bugaboo--an imaginary terror--caused more mischief and delay in the next dozen years than any other. Further, it was fuzzy in that it seemed to mix or confuse the concepts or requirements of including minorities in a validation study and performing separate validation studies for minorities. Of course when validation studies were referred to, the writers only meant "criterion-related validation studies." Anything other than criterion-related validity seemed beyond the authors of these 1966 guidelines.

Nowhere was the real issue frontally addressed, i.e., the oft-noted differences in test scores between minorities and others. This omission is most curious since it had been resurrected in the mid-1960's from the early 1950's, and was a source of major interest in the late 1960's as a result of some rather inconclusive studies. In 1967 the technical spokesperson for the EEOC in Congressional testimony described the results of two of these studies of the "new concept, differential validity" as "truly amazing" with implications that could be enormous. In 1969 one more study was added and the representative concluded"...much evidence has been accumulated that minorities' test scores may underestimate their job performance..."

It was then only a small step to include a requirement for differential validation in the 1970 EEOC Guidelines and in the 1971 Department of Labor Order.

Before warning the public about the hazards of cigarette smoking, the Surgeon General responsibly developed comprehensive, publically reviewable and reasonably convincing evidence available to the Federal government regarding differential validity or differential prediction in 1970. Nevertheless, based upon an untested hypothesis, test users were put on notice, clearly without meaningful scientific support, that it would be unacceptable to use a test absent such a study. The 1970 Guidelines were subsequently disavowed by the advisory committee which had helped work on an earlier version, but it took four years for the committee to state publically,"...these published Guidelines contained material which had never been seen in any form by members of the advisory committee and with which most members took great exception as being either untenable or unworkable..." But, absent a meaningful challenge, the concept became entrenched and the root cause of years of inter-agency acrimony and wrangling on this and a number of issues which metastasized from it.

In the meantime, other researchers had begun some serious study of the issue. In 1966 the Educational Testing Service and the [then] U.S. Civil Service Commission began a cooperative research effort to study the fairness to Blacks and Chicanos of a variety of employment tests for different kinds of occupations. The study was to take six years. The results demolished the viability of the concept and caused responsible professionals to rethink and, in some cases, to renounce their prior views. There came an electrifying day in June 1972 when the results were reviewed and commented upon publically in a forum which I co-chaired. Here are some of the words of Bob Guion at that meeting:

"...In light of my previously published views, the findings of these studies are not personally very satisfying...I would summarize the information here, and that emerging in the general literature as well, by suggesting that, as a general rule, the validity of a test against a specified criterion is likely to be about the same for all comers..."

And the late S. Rains Wallace at the same public conference:

"...It appears to me to be about time for us to accept the proposition that written aptitude tests, administered correctly and evaluated against reasonably reliable, unbiased, and relevant criteria, do about the same job in one ethnic group as in another.

"It seems clear that people like me who expected to act as a moderator variable for validity relationships were wrong. It also seems clear that people who assumed that all written tests were inappropriate and unfair instruments if applied outside of the WASP culture were equally wrong..."

Thus, in the views of many leading responsible professionals, the issue of differential prediction/test fairness was reasonably resolved by mid 1972.

The now defunct Equal Employment Opportunity Coordinating Council, established by law in March 1972 held its first meeting in November of that year and directed its attention to the testing issue and to the desirability of a common federal agency position on testing guidelines. (There were then three in existence: EEOC, OFCC, and [illegible].) Staff from the involved five EEOCC agencies met several times in November and December 1972 and in January 1973. The staff group had been directed to get together and set out their differences on testing, and so did on January 31, 1973. On February 8, 1973, the principals of the EEOCC directed the staff group to reassemble, iron out the differences and, within a month, produce uniform testing guidelines. This directive derived from the intent of Congress and the will of the President, yet it took over five years, closer to six, three presidents, dozens of federal officials and the time of hundreds of commentators before that directive was complied with.

Following the February 8, 1973 directive, the chief staff representatives from each agency agreed upon 13 principles which was to make the guideline writing process easier (February 26, 1973). Two bear looking at: the good news and the bad news, so to speak.

The good news was that there would not be a preference among the three validation strategies; the choice would depend upon the situation. EEOC staff was aghast. Its position in active litigation was undermined as well as its posture in regard to all employers that criterion-related validation was the preferred method and that resorting to any other method required a prior proof that criterion-related validation was infeasible. In defiance of their own staff director, EEOC attorneys wrote to the Department of Justice on March 27, 1973, "...The policy which the Commission [EEOC] has followed is that an employer acts at his own peril when he uses some other approach..." (Note that EEOC viewed employers in the masculine gender.)

But "...at his own PERIL..."! If that wasn't government arrogance at its worst. Imagine how many of us were living in a perilous world, unbeknown to EEOC. Despite this agreement of principle, as late as the Spring of 1978 the Justice Department attacked an employer's use of content validity on the grounds that it presented no risk since, if followed properly, the procedures could always lead to a conclusion of validity. Criterion-related validity, however, presented risks, the Department said, and therefore that was why the employer did not use it!

Well, all that was the good news. Now, the bad. Differential prediction would be included in the Guidelines. I was appalled. After six years of research in which I had been personally involved, the expenditure of millions of dollars--a lot of it federal money--and the resultant conclusion that this was a non-phenomenon identified erroneously by some badly conducted studies, here it was alive, well, and being fertilized. Justice and the EEOC simply ignored the mainstream findings of professional research although bombarded over the next five years by objections from the psychological profession.

I suppose that I have contemplated this curiosity more than any other because its inclusion spread pernicious roots into other aspects of the Guidelines as well as the developmental process itself. If it had been abandoned, however, you can imagine how EEOC personnel credibility would have been eroded.

I have written elsewhere that in the Guidelines developmental process there were no tradeoffs. I had meant "development" to encompass the writing process. While I have never discussed with them what went on among the agency staff chiefs, I have a strong feeling that there was a tradeoff of the parity of the validation strategies for the inclusion of differential prediction. It was a poor tradeoff, because I believe we would have won the first anyhow based upon professional consensus.

Even as late as the Spring of 1978 Jim Scharf, when he was with EEOC attempted to get reconsideration of inclusion of this section in the Guidelines. This was a view that he had shared with me as early as 1976: that the continued inclusion of differential prediction would badly serve the groups that his agency was interested in. But it was to no avail. After all these assaults, a Justice Department attorney who had been in the thick of things since 1972 and had urged that it (differential prediction) be allowed to "go on a little more," opined, with a straight face, in 1978, that "I understand some people don't believe it exists."

The assertion through the Guidelines that differential prediction was alive and well in the face of the overwhelming evidence to the contrary was summarized by Frank Schmidt as follows: that the refusal to accept scientific findings, as it has been through the ages, holds firm because it contradicts deeply held social, political, or religious beliefs. He aptly illustrated this with the reactions of the wife of the Archbishop of Canterbury upon hearing for the first time about the theory of evolution. Her statement was: "It's not true, and if it is, let us hope it does not become generally known."

I opine (with a probability of being right greater than .95) that the requirement was left standing because certain agencies of the federal government anticipated that they could tolerate negative professional reaction more than negative reaction from other constituencies.

In a rare show of deference to professional standards, federal officials excused the inclusion of differential prediction on the grounds that the 1974 APA Standards required such an investigation. What a chicken and egg situation! The 1974 requirement was there because of "...regulations pursuant to civil rights legislation..." At any rate, I note that it is downgraded in the 1985 APA Standards as a requirement.

In the end, of course, allowing it to stand was and still is a blatant insult to the members of and to the very groups about which the Guidelines were concerned. Minorities and women are capable of facing facts, but agencies charged to be their advocates shielded them from the true state of affairs. In the first place, to hold out any hope that these studies would be done in abundance was to fly in the face of the Guidelines themselves. The Guidelines allow the user to determine if a criterion-related validation study is feasible. Schmidt, Hunter, and Urry came to the rescue on this issue ten years ago with their seminal article on sample size. If users follow their guidance, the N's required as a practical matter are beyond most employers. Studies might be done by large wealthy employers, groups of employers, or by test publishers who might be able to assemble large N's across organizations. The Schmidt et. al. article contributed to our tacit decision not to rock the boat any more and to leave differential prediction in the Guidelines since such studies were, as a practical matter, virtually impossible to do. But as long as the requirement is left standing there is a misleading signal being sent by the federal government that somehow it believes tests do behave differently for different groups.

If few employers can do meaningful criterion-related studies, where does that leave us except in the arms of content and construct validity? Once again, Schmidt and Hunter to the rescue. They have spent years assembling the data so that we really need not be on the treadmill of criterion-related studies. The validity generalization work done by Schmidt, Hunter, and their followers may just be the most important original measurement contribution of the last decade. It is seldom that one sees one's work acknowledged in the professional measurement Standards in one's lifetime. But the 1985 Standards, I note, recognize this development and offer guidance for transportability.

During the development of the Guidelines a number of other issues surfaced which were susceptible of resolution through reference to and deference to the existing research literature but which were intuited along for several years. For example, one of the enforcement agencies continually worked to establish minimum cutoff test scores for many jobs (truck driver was often cited) because, it was alleged, that more of the skills would logically not make for better job performance. I listened to this for a long time and then one day Hawk from the Department of Labor came to my office with a copy of his study of the linearly of some 17,000 regression coefficients.

The results clearly showed that for a wide variety of tests and jobs non-linearity was a chance phenomenon except for certain types of personality measures which are not at issue. Thus the basis for both cutoffs and ranking throughout the range of scores is not only permissible but supportable.

The point which emerges is that the development process was basically flawed in that a scientific basis for the Guidelines was always a grudging last resort. Where scientific knowledge is established such as in differential prediction, linearity of regression, etc., the burden should not be upon a test user, but should shift to those who claim inappropriateness for a specific situation. Science, more often than not, got in the way of what the federal enforcement agencies wanted to do. I do not mean to imply venality; rather, I believe the actions can be attributed to the elan that typically fuels new governmental initiatives. At the same time, please be aware that case law was written into the Guidelines as fast as it developed. As a result the Guidelines is primarily a litigating document. In the waning days of its construction the Office of Personnel Management withdrew and withheld its active legal involvement leaving lawyers from Justice to EEOC to do exactly what they wanted under the blessing and protection of a political leadership committed to numbers not merit; to intuition, not knowledge; and to onerous employer burdens in the belief that it would be easier to hire minorities and women than to meet the technical requirements of the Guidelines. Please be aware that the Justice Department which, in the Spring of 1978, had courted the APA Committee on Tests and Assessment to try to secure an endorsement of a late draft of the Guidelines did not even bother to try for that endorsement for the final Guidelines. Either this was a case of "Don't ask the question if you don't want to hear the answer," or the arrogant confidence that the federal government no longer needed such an endorsement.

In the end I signed off on a recommendation to publish the Guidelines in the belief that they were consistent with the policy goals of that particular administration. Please be aware that I never signed the Guidelines themselves. I did speak to groups about the Guidelines, made a couple of training films and was generally encouraging and in a position to see that they got implemented in the federal government. These activities were the loci of my last months at OPM and my farewell, my valediction, to the Guidelines themselves. I have not spoken or written publically about them until now. I doubt that I will again. I do not intend, like many famed concert stars, to give lifelong "farewell tours," at least not singing the same tunes.

But in seven years absent the scene I have, as I suggested earlier, acquired certain perspectives which I want to share with you. First, not only was the developmental process flawed but the basic assumptions themselves were flawed. I would suggest that those who believed that tough testing guidelines would cause a significant increase in the employment of minorities and women were wrong. Employers got smart: they learned how to validate tests. Ms. Norton, once Chair of the EEOC, and certainly one of the most able persons to hold that office said, in the 1970's:

"My hat is off to the psychologists." She did not see "evidence that validated tests have in fact gotten black and brown bodies, or for that matter females into places as a result of the validation of those tests. We do not quite see the causal relationship we had expected to see."

What non-psychologists failed to anticipate is that, despite Guidelines and the onerous documentation requirements, while the process of validation may seem mystical and difficult, once the rudiments are competently carried out, validity is darned difficult to avoid.

Second, let's put the Guidelines in perspective. They are not the center of the personnel measurement universe, and that is most probably why I did not deal with them ten years ago. The history that I have touched upon today spans mostly a 14 year period, 1972-1986. It seems only a moment ago that we began. In another of those moments it will be the year 2000. Over 26 million new jobs will become available by then in the United States. But there is a University of Chicago study which projects that by that time Black male employment will fall to 30%! The presence or absence of testing guidelines will have nothing to do with this unfolding American tragedy. Certainly this issue eclipses in importance most others, and it is virtually unrelated to equal employment opportunity both in etiology and in solution.

The decade of the 1970's will be remembered as a defensive reactive one in and for public personnel management. The only game in town, and the cornerstone of public policy concern in most matters was equal employment opportunity. When I ceased active involvement in professional activities seven years ago, it was difficult to find a measurement conference, a professional meeting, seminar, training session, workshop, symposium, or what have you that was not primarily centered around EEO. We wanted to respond to this nationally important agenda item. In our responses we addressed (perhaps some for the first time) measurement and other personnel practices which had been fundamental to public personnel management for so long that they seemed almost sacrosanct.

Each of us has had the experience of, when reading a book, turning the page and finding that what we're reading doesn't seem to connect with what we've just read. We have simply turned more than one page. I last looked at this organization and what is was doing in 1979. I look again today and find I have not turned one page too many but I am into a new chapter or almost another book. You have moved from reactive to proactive; from defensiveness to assuredness; from past to future. The evidence is in the list of publications in a recent issue of PSYSCAN; from the IPMAAC News; and in the subjects of this and prior annual meetings and workshops. I have reviewed these many times. My content analysis clearly shows that you are in a new chapter of professional excitement, development, diversification and progress, the intensity and character of which has rarely been matched. This is what we hoped for in 1976; scientific and process development and improvements in our field and the use of them.

Today I have said the infusion of professional scientific knowledge into the Guidelines was grudging when it should have been the only sound foundation for their development, issuance, and enforcement. My view was and is that we do not need technical guidelines beyond professional Standards so long as the latter are kept current with the state of knowledge.

Having said that, I am going to contradict myself It was probably better to have had federal guidelines, even with their flaws, than not to have had them. I believe they helped stimulate the very advancements and improvements that I see and am commending today. Perhaps these would have come about anyway, but I believe the Guidelines stimulated them both in speed and content. Having gained a momentum of their own, professional advancements and contributions are now being made independent of and virtually without reference to, federal testing guidelines. They have a life of their own, thanks to those of you who have made contributions, who have used them, and who have made this organization an important professional resource and conduit.

While I said farewell to Guidelines seven years ago and left the scene, that is not the valediction which is the one I honor today. YOU have met the Guidelines and, in one way or another, conquered them. YOU have gone so far beyond them that it is YOU who have said farewell to them. YOU have said a far more meaningful farewell by making them an obsolete curiosity of the past. YOUR work has made YOU, individually and collectively, the valedictorians, the winners, and the charters of the future. Your professionalism today makes me rejoice to have been a part of your early chapters. I thank YOU for enriching my life and for asking me here today.

* * *

ASSESSMENT CENTER TOPICS (Paper Session)

The Assessment Center: Effects of Pooling on Dimension-Specific Ratings

Phillip E. Lowry and Clinton Richards, University of Nevada Las Vegas

One of the principal concerns of personnel administrators is the development and selection of personnel. The assessment center is an important tool for these joint purposes. Properly conducted assessment centers have been shown to be reliable predictors of job success and appropriate for affirmative action programs (e.g., Howard, 1974). The assessment center is, however costly in comparison with many other commonly used techniques for personnel selection.

The pooling of assessor judgments is one practice that adds significantly to the cost of an assessment center. If candidate ratings could be determined by a simple arithmetic decision rule based on independent assessor judgments without significantly savings could be realized. However, the most definitive current guidelines for the assessment center process clearly support the pooling of assessor judgments. According to the Standards and Ethical Considerations for Assessment Center Operations, (Task Force, 1980), judgments should be "pooled by the assessors at an evaluation meeting during which assessment data are reported and discussed, and the assessors agree on the evaluation of the dimensions and any overall evaluation that is made."

The focus question of the present study is whether dimension-specific pooling has a significant impact on ratings. Several practitioners have previously reported that the overall ratings obtained by pooling were highly correlated with overall ratings obtained by arithmetic rules only (Russell, 1983, Joiner and Carlin, 1983, 1985). However, Sackett and Wilson (1982) found less disagreement among assessors (prior to pooling discussions) on overall ratings than on dimension-specific ratings.

METHOD

Data for this study were collected during three assessment centers conducted for city governments. Two of the three assessment centers were selection centers. One was a career development center. Fourteen individuals were rated on five dimensions by thirteen assessors. Two scores were developed by the assessors for each participant on each dimension; the prepooling score (the raw arithmetic score before any discussion), and the consensus score (the agreed upon score after discussion). Our primary hypothesis is that the performance dimension scores will be significantly changed by the pooling discussions.

Multivariate analysis of variance (MANOVA) (Hull and Nie, 1981) was used to examine the effects of pooling on candidate scores on 5 performance dimensions. The results of the MANOVA analysis indicate a significant pooling impact (approximate F of 2.67, significant at .03 level). All but one of the performance dimensions, written communications, was significantly changed from the pre-pooling to consensus rating periods. Scores on oral communications, problem solving, decisiveness, and influence all changed significantly. Scores changed more in the development center than in the selection centers. Changes were great enough to affect participant rankings only in the development center.

DISCUSSION

This research suggests that performance dimension scores do change significantly as a result of the pooling process used in this study. The results are particularly important for those who use assessment center scores together with other criteria for making selection decisions. In this case, the changes in performance dimension scores can have a significant and important effect on the total standing of a participant even if the rankings

remain the same before and after the pooling discussions. The impact may be even more pronounced when the selection authority differentially weighs the dimension scores. In development centers, even small changes in performance dimension scores could have an impact on feedback given to participants. In fact, the information available for feedback could increase as a result of pooling even though dimension scores did not change.

More research is needed on the impact of pooling. A number of conditions may influence the observed effects. First, variations in the pooling process itself may produce different effects. Our research indicates that dimension specific pooling does have a significant effect on scores. The research of Russell (1983) and Joiner and Carlin (1983, 1985) suggests that pooling for overall scores has only a small effect.

Pre-pooling evaluation procedures may also influence the effects of pooling. For example, pooling is likely to have less impact when assessors are able to observe all participants in all exercises. This was done in the two selection assessment centers but not in the development center. Perhaps this explains in part the apparently greater impact of pooling in the development center. In interviews conducted after the development center was concluded, assessors were emphatic about their felt need for pooling.

Another difference between the development and selection centers which may have affected the results was the differences among assessors. In the selection centers the assessors were essentially homogeneous with respect to their job background and culture. They were fire service officers assessing fire service officers. On the other hand, in the development center the assessors had completely different job backgrounds, from each other and from the participants.

REFERENCES

Howard, A. "An Assessment of Assessment Centers," Academy of Management Journal, 17 (January 1974), 1150134.

Hull, C.H. and Nie, N. SPSS Update, Versions 7-9. New York: McGraw-Hill, 1981.

Joiner, D.A., and Carlin, P. "Consultant-Agency Cooperation in Conducting Research on a Promotional Assessment Center for Police Lieutenant." Proceedings of the 1983 IPMA Assessment Council Conference on Public Personnel Assessment, pp. 39-40.

Joiner, D.A., and Carlin, P. "Further Research on Assessment Centers", 1985 IPMAAC Conference, New Orleans, contained in tape number 19, Convenient Cassette Service, P.O. Box 6931, Metairie, LA 70009.

Russell, C.J. "An Examination of Internal Assessment Center Processes for Compliance with the Uniform Guidelines." Proceedings of the 1983 IPMA Assessment Council Conference on Public Personnel Assessment, pp. 38 39.

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

Task Force on Assessment Center Standards. Standards and Ethical Considerations for Assessment Center Operations. The Personnel Administrator, 1980, 25(2).

* * *

Professional and Legal Standards Related to Assessor Training for the Assessment Center Method

Patrick T. Maher
Personnel & Organizational Development Consultants, Inc., La Palma, CA

Assessment centers require trained assessors, but many public agency assessment centers do not use adequately trained assessors. Fitzgerald and Quaintance (1982) voiced some concerns of assessor training time reported by some jurisdictions of from one hour to five days. Yeager (1986) and Byham (1977) also found ranges of no training to three weeks of training.

Assessment centers are covered by the Standards and Ethical Considerations for Assessment Center Operations (Standards), first issued in 1975, and revised in 1978 with an expansion of the section dealing with assessment center training and guidelines to determine assessor competence.

It is well recognized that length of training is not relevant to quality or relevance of training. The Standards envision an assessor certification program which could ensure the adequacy of training and the adequacy of learning.

Evidence exists, however, to support the idea that adequate assessor training will require a minimum amount of time. Jaffee (1985), Brademas (1985), Humphreys (1986), Maher (1984), and Byham (1977) state that assessor training requires three to five days.

In examining the assessor's role and its relationship to the validity of assessment centers, Olshfski and Cunningham (1986) found assessor training to be especially important, particularly as it is applied to careful observation and thoughtful attention to judgments based on observed behavior.

Research for this paper indicated that, while assessor training has been virtually ignored, other relevant information exists. For example, Wexley, Sanders, and Yukl (1973) found that contrast effects can only be reduced by a fairly-intensive training program. Latham, Wexley, and Pursell (1975) found that performance-measurement variance due to rater differences can be reduced by training observers to minimize rating errors. Ivancevich's (1979) research findings support other research on the importance of training effects in reducing psychometric error, and showed that intense training significantly reduced halo and leniency error.

Just as there is a lack of professional literature dealing specifically with assessor training, there is a similar vacuum in the legal issue. Byham (1980) reviewed "all known court cases dealing with the assessment center method" as of January 1, 1980. The majority of cases involving assessment centers seem to be resolved on issues other than the adequacy of the assessment center process itself.

In the first case involving the legal adequacy of the assessment center, the often-cited Berry v. City of Omaha, a variety of issues were raised, including whether assessor training was adequate. The court found that adequate and comparable training allows different groups of candidates to be fairly assessed by different groups of assessors.

The only other major case that deals with the issue of assessor training is Fire v. City of St. Louis. The city's validation report anticipated at least three to four days of training to "assure standardization of assessment." In the actual administration of the assessment center, only two days of training was given for interview and training simulations, and one day of training for those assessing a fire simulation. The appellate court found the raw data showed substantial variance among the ratings given by the assessors in that the statistical coefficients of correlation gave an incomplete picture of the reliability of the procedure. While the appellate court overturned the district court's finding that the fire simulator was a job-related examination procedure, they were hesitant to hold that the district court erred in sustaining the validity of the interview and training portions of the assessment center procedure.

Some have advanced the idea that assessor training must be assessment center specific. This concept does not appear to have been addressed in any of the literature, and a critical examination of it would tend to refute it.

In addition to adequate training, there is a need for certification of trained assessors. Frank and Whipple (1978) report that there is an obvious need for the development of a comprehensive assessor certification program. Cohen (1978) states that one way to reduce the likelihood of vast assessor ability differences is to implement a certification procedure required of all assessors after training but prior to actual assessment duties.

The extent to which assessor training, length, as well as content, impacts the reliability of the subjective decisions of assessors should be addressed. In addition, the importance of assessor training in establishing consistency of grading scores between various groups of assessors and/or candidate groups must also be addressed, especially given the potential that these specific issues may be raised in later legal challenges. (See, by way of example, David v. Michigan Civil Service Commission.)

While the courts have not dealt with the training of assessors to any great extent, it is likely that we can anticipate court challenges in the future.

Since the Standards place extremely strong emphasis on trained assessors, it seems that considerably more attention would have been devoted to this particular aspect of the assessment center procedure.

REFERENCES

Brademas, J. Personnel Correspondence, May 1, 1985.

Burke, M.J. & Langlois, G.M. Assessor Training: A Review of the Literature and Current Practices. Journal of Assessment Center Technology, 1981, 4, 1-8.

Byham, W.C. Review of Legal Cases and Opinions Dealing with Assessment Centers and Content Validity. Pittsburgh, PA: Development Dimensions International, 1979.

Byham, W.C. Assessor Selection and Training. In J.L. Moses and W.C. Byham (Eds.) Applying the Assessment Center Method. New York: Pergamon Press, 1977.

Cohen, S.L. Standardization of Assessment Center Technology: Some Critical Concerns. Journal of Assessment Center Technology, 1978, 1, 1-10.

Fitzgerald, L.F. & Quaintance, M.K. Survey of Assessment Center Use in State and Local Government. Journal of Assessment Center Technology, 1982, 5, 9-22.

Frank, F.D. & Whipple, D. An Assessor Certification Program Based on Simulation of the Assessor Job. Journal of Assessment Center Technology, 1978, 1, 1-14.

Ivancevich, J.M. Longitudinal Study of the Effects of Rater Training on Psychometric Error in Ratings. Journal of Applied Psychology, 1979, 64, 502-508.

Jaffee, C.L. Assessment Centers: Present and Future Perspectives. WRIPAC Invited Speaker at the International Personnel Management Association Assessment Council 1985 Annual Conference.

Latham, G.P., Wexley, K.N. & Pursell, E.D. Training Managers to Minimize Rating Errors in the Observation of Behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Maher, P.T. Assessor Training Manual for Public Sector Assessment Centers. La Palma: Personnel & Organization Development Consultants, Inc., 1984.

Maher, P.T. An Analysis of Common Assessment Center Dimensions. Journal of Assessment Center Technology, 1983, 6, 9-22.

Olshfski, D.F. & Cunningham, R.B. Establishing Assessment Center Validity: An Examination of Methodological and Theoretical Issues. Public Personnel Management, 1986, 15, 85-98.

Standards and Ethical Considerations for Assessment Center Operations, 1978, Task force on Assessment Center Standards.

Standards and Ethical Considerations for Assessment Center Operations, 1975, Task force on Assessment Center Standards.

Wexley, K.N., Yukl, G.A. & Kovacs, S.Z. Importance of Contrast Effects in Employment Interviews. Journal of applied Psychology, 1972, 56, 45-48.

Yeager, S.J. Use of Assessment Centers by Metropolitan Fire Departments in North America. Public Personnel Management, 1986, 15, 51-64.

Zedeck, S. Performance Measures: Forms or Samples? Summary of an invited talk at the International Personnel Management Association Assessment Council 1984 Annual Conference.

CASES

Berry v. City of Omaha, Douglas County, Nebraska District Court, November 17, 1975.

Davis v. Michigan Civil Service Commission, Ingham County, Michigan, Circuit Court, 78-21743-AZ, June 16, 1978.

FIRE v. City of St. Louis, 616 F.2d 350 (1980).

* * *

Defending Your Assessment Center Against the Experts: A Case Study

Richard C. Joines

Management & Personnel Systems, Inc., San Francisco, CA

Introduction: In 1982, the author developed a promotional examination for Battalion Chiefs in the San Francisco Fire Department. The examination consisted of a multiple-choice test, two leaderless groups discussion (LGD) exercises and an individual problem analysis/report exercise. There were 82 candidates. After the examination, a group of candidates, primarily consisting of individuals who had been operating as Battalion Chiefs on temporary appointments, filed a case against the examination in Superior Court (Carrozzi et. al. v. Civil Service Commission of the City and County of San Francisco, California Superior Court No. 805-940, 1983). The City/County prevailed in defending the examination.

The case was decided based upon the administrative record—which consisted of reports filed by three psychologists retained as experts by the plaintiffs, the author's report in defense of the exam and a transcript of a six hour hearing before the San Francisco Civil Service Commission. The Civil Service Commission hearing and the reports filed by the experts for both sides focused upon the reliability and validity of the assessment exercises. This paper reviewed the arguments set forth by the experts for the plaintiffs coupled with the ways in which these arguments were rebutted by the author. The summary which follows addresses the more significant technical issues that were addressed.

Issue: Weighting the Exam Parts

Opposing Experts: The announcement for the examination stated that 1000 points would be possible, as follows: multiple choice test = 550; two LGD exercises = 170; report exercise = 200; seniority = 80. The opposing experts argued that the effective weights of the exam parts were not equivalent to the announced weights. Due to a larger standard deviation, the assessment portion carried a greater weight in determining overall rank-ordering on the list. The actual weight of the multiple-choice test was 44%, not 55%; and the actual weight of the assessment portion was 48%, not 37%. Within the assessment portion, the written report carried an effective weight of 20%, not 17%; and the LGD exercises had an effective weight of 17%, not 20%.

Rebuttal: The examination did not list percentage weights for any of the test components, but rather, total points possible on each component in an examination that had 1000 points possible. Thus, the Commission had not specified percentage weights that the component test parts should carry. Moreover, standardization of scores in this case would have had minimal impact on the ranking of candidates.

Issue: Consistency between Written and Oral Exercises

Opposing Experts: Argued that the correlations between common assessment dimensions in the report exercise and the LGD's should have been higher. The assessment dimension, judgment & decision making, was rated in both the report exercise and the LGD's. The correlation was only .10, whereas the correlations of the assessment dimension within the written exercise and within the LGD's was significantly higher. This suggests that the ratings within exercises were largely a function of halo and that the dimension themselves were meaningless.

Rebuttal: The correlation reported by opposing experts was incorrectly calculated. In actuality, the correlation for judgment and decision making between the report and LGD exercises was .49. This correlation is reasonable and consistent with other reported research.

A general relationship between scores on common dimension assessment dimensions between written and oral assessment exercises would be expected. However, there is no necessary degree of correspondence required in order to support the validity of the separate exercises. The written and oral assessment exercises are not designed or intended to produce correlations comparable to those obtained for parallel forms of a test; if this were so, there would be no need to use both written and oral exercises.

Differences in candidate scores are expected. The written exercise required analysis of a number of administrative and fire related issues on an individual basis, coupled with the ability to commit the analysis to a written report. The problem solving and decision making skills elicited by the LGD required the ability to incorporate the ideas and points made by others as well as convey information in an understandable and cogent manner. Thus, both written and oral exercises are included in the process and the obtained correlations were reasonable. They were not indicative of deficiencies in the validity of the process as charged by opposing experts.

Issue: Assessor Training

Opposing Experts: Argued that the length of the assessor training program was insufficient. They maintained that a good program would be on the order of three weeks, with five days being a bare minimum. References to some private sector training programs were made in support of their argument.

Rebuttal: There is no consensus within the profession on the length of training time required for assessors to function properly. Some experts believe that the training program should be at least equivalent to the length of the assessment process, whereas others believe it should be double this amount.

Approximately one day of training was provided the assessors who rated the report exercise. Three and one-half days were devoted to rating 82 reports. All of the assessors were one management level higher than the candidates and were from major fire departments. The assessors understood the problems

contained in the exercise, were trained in the three assessment dimensions that were rated, and were provided standardized guidance for in the form of points to consider in reviewing candidate reports. Thus, one day of training was provided for a 90 minute exercise, with only three dimensions being rated. This is not comparable to the private sector training programs used as comparisons—programs which may assess candidates from three to five days using multiple assessment formats and rating candidates on 10-20 dimensions.

Assessors were also provided one day of training in observing and evaluating candidate performance in LGD's. Two LGD's were used. Assessors had sufficient time to review the LGD problems. Three dimensions were evaluated in the LGD's and each dimension was well-defined and anchored with positive and negative behavioral examples. The assessors were trained in observing behavior, classifying behavior and evaluating behavior. They completed practice exercises and observed an LGD videotape of fire personnel.

Issue: Choice of Assessors

Opposing Experts: Argued that college professors or professional psychologists should have been used.

Rebuttal: Research supports the view that managers one level higher than the target position can function just as effectively as assessors and psychologists or others.

Issue: Test Security

Opposing Experts: Argued that use of the same LGD problems over a period of four days compromised the exam and benefitted those candidates who took the exam later in the week. In response to the rebuttal position that there were no significant differences between the mean scores for any two days during the four day exam period, the opposing experts argued that this could be explained by assessors raising their rating standards to offset the advantage of the candidates who reported later in the week. In effect, the argument was that the raters simply fit their ratings to a bell curve, thereby penalizing candidates who happened to be in LGD groups with exceptionally skilled individuals or individuals whose performance was better as a result of foreknowledge of the LGD problems.

Rebuttal: In addition to establishing that there were no significant differences in mean scores across the four days the LGD's were administered, raters were required to base their ratings on preestablished behavioral criteria. The assessors were trained to assign a positive rating to an individual who demonstrated behaviors considered to be positive on an assessment dimension—regardless of the level of competition or behaviors of other individuals. Thus, candidates were rated against external criteria. All candidates in any given group of five or six LGD participants could have scored high; or all could have scored low. Thus, opposing experts were wrong in their contention that candidate ratings were simply fit to a bell curve.

Issue: Size of LGD Groups

Opposing Experts: Argued that is was improper to have some LGD groups consisting of six candidates, whereas in most instances there were only five participants. Given a 60 minute LGD with five participants, each participant would have an average of 12 minutes of active participation. With six candidates per group, only 10 minutes would be available. Facing this discrepancy, bogus candidates should have been used (role players) to form groups consisting of six candidates across all LGD's.

Rebuttal: Given 82 candidates, it was necessary to have 14 LGD's with five candidates and two LGD's with six candidates. The sample of behavior available in the six candidate groups was not substantially different from that available in the five candidate groups. Using bogus candidates across fourteen LGD's is no solution at all. Two groups would not have been standardized. Bogus candidates might vary their behavior from one group to the next, further lessening standardization of the process.

Summary: This case involved a number of technical issues relevant to the way in which assessment centers in the public sector are conducted. Space does not permit coverage of all these issues; hopefully those which have been described will offer meaningful insights into the kinds of issues that may become the subject of litigation.

* * *

Employee Drug and Alcohol Abuse - Industry's Approach

Peter P. Greaney, M.D., University of California at Irvine

The annual cost to industry of employee drug and alcohol use has been estimated at up to $16.4 billion dollars. A confidential mail survey of national organizations conducted in 1981 reported that 80% of the respondents had to deal directly with drug problems (1). While alcohol was the most commonly abused substance (82%), marijuana incidents occurred in more than half the firms (55%), and both heroin and cocaine use reported by one-fifth of the organization. The survey confirmed that drug usage in the workplace is relatively widespread and it is not confined to blue collar minority groups. An employee whose drug and alcohol usage impairs his or her health and interferes with safe efficient work performance has a problem. Irrespective of whether the employee uses a drug off or on the job, or even the type of drug used, the behavior induced by drug use reduces employee performance, lowers employee morale and increases the risk of accidents.

Employers use a variety of means to combat employee drug and alcohol use. The most widely used technique is to develop a company policy on alcohol and other drugs. Policy manuals outline the organization's position on drug and alcohol abuse, including acute drug intoxication on the job and the buying and selling of illicit drugs at the workplace. Another procedure is to establish an occupational treatment program whose primary target is workers whose job performance is impaired. The employee assistance, troubled worker or broad brush approach to the issue of employee substance abuse has proven viable in many business settings (2). The employee assistance program (EAP) is a confidential service that intervenes with troubled workers, whether self or supervisor referred, and provides training to supervisors, union representatives and employees. Intervention varies with the particular program from simple triage to diagnostic evaluation, motivation, referral and follow-up. The treatment, normally subsidized by the company but provided at an accredited treatment facility not affiliated with the firm, usually is considered a condition of continued employment. It is estimated that 50-75% of all EAP referrals involved alcohol misuse, and rehabilitation rates average 70% of referrals.

Although the weight of evidence suggests that occupational programs are relatively effective, current limitations reduce their overall effectiveness in maintaining a drug-free work force. Most EAP programs reach only 5% of the target population and case finding methods in a majority of EAP's are crude. Monitoring program success is difficult as success has been defined in various ways from significant improvement in job performance to modified drinking/drug taking behavior. The traditional program assumes that an employee's value to the organization is based on substantial training and time investment, a value that often does not extend to the youthful abuser. Young employees, having a different work ethic, do not respond favorably to constructive confrontation.

Another method of providing a drug free work place involves urine drug screening on all employment applicants and selective screening of suspected abusers. Urine toxicology screening is an effective test to determine the presence of drugs in the urine. Thin-layer chromatography and radio immune assays or modified techniques, such as enzyme multiplied immuno assay technique (EMIT), test for a wide spectrum of drugs including marijuana, PCP, heroin, opiates, amphetamines. A great deal of weight is often placed on positive findings; however, the test does not provide information about the pattern of use and cannot distinguish between the occasional user and the chronic abuser (3). The cost effectiveness of this approach can be improved by limiting the tests based on the results of the pre-employment medical examination (4). There are serious questions about the reliability of the results of screening urine for drugs. In a recent evaluation of the performance of 13 laboratories, error rates for amphetamines, barbiturates, methadone, cocaine, codeine and morphine ranged from 11% - 94%, 19% - 100%, 0% - 33%, 0% - 100%, and 5% - 100%, respectively. False positives ranged from 0% - 6%, 0% - 37%, 0% - 66%, 0% - 6%, 0% - 7%, and 0% - 10%, respectively (5). The results suggest the need for monitoring the performance of organizations and contract laboratories with blind quality-control samples.

All urine toxicology screening tests require confirmation by a alternative method prior to being considered positive. Where punitive action is contemplated, additional tests may be necessary to accurately quantify urine and serum drug levels.

The problems of drug use among employees is steadily increasing and has not been thoroughly investigated. As none of the above three approaches to maintaining a drug-free workplace is ideal, organizations may wish to consider using a combination of policy development, pre-employment examination with selective urine toxicology screening, employee education, EAP referral and rehabilitation, recognizing the limitations of each methodology.

REFERENCES

1. Schreier, James W., A Survey of Drug Abuse in Organizations, Personnel Journal, 478-485, June 1983.

2. DuPont, Robert L., M.D. & Basen, Michele, M., M.P.A., Control of Alcohol and Drug Abuse in Industry - A Literature Review, Public Health Reports, Vol. 95, No. 2, 137-148, March-April 1980.

3. McBay, Arthur; Dubowski, Kurt; & Finkle, Bryan, Urine Testing for Marijuana Use, JAMA, Vol. 249, No. 7, 881, February 18, 1983.

4. Lewy, Robert, Pre-Employment Qualitative Urine Toxicology Screening, Journal of Occupational Medicine, Vol. 25, No. 8, 579-580, August 1983.

5. Hansen, Hugh; Caudill, Samuel; & Boone, Joe, Crisis of Drug Testing, JAMA, Vol. 253, No. 16, 2382-2387, April 26, 1985.

* * *

INNOVATIONS RELATED TO WORK SAMPLES, SIMULATIONS, AND IN-BASKETS

Clerical Work Samples: Three Practical Approaches to Scoring

Janet L. McGuire, Psychological Services, Inc., Washington, D.C.

INTRODUCTION

Staffing and testing specialists in State and local governments face a variety of practical problems in developing tailored work sample tests for entry and promotional clerical vacancies. Job content can vary widely

across both occupations and specific vacancies, and the professional literature offers little in the way of guidance on how to adapt methods and procedures from large-scale, standardized testing to smaller-scale, tailored applications.

Some jurisdictions are able to sidestep the difficult problem of how to award points and set cutoff scores by the use of creative crediting and certification approaches. Among these are banding, overall judgment of "qualified" vs. "unqualified", and other approaches. Other jurisdictions must meet rigid civil service rules requiring 70% pass points, elaborate tie-breaking procedures, and the like.

For many jurisdictions, however, the process of developing a scoring approach and determining reasonable passing points is a tortuous one. This presentation describes three testing situations where there was a need for an understandable and rational explanation for both the scoring approach and the cutoff score. These approaches were developed in a small local government setting for use in filling individual clerical vacancies with a high degree of political sensitivity or clerical union interest.

CLERICAL POSITIONS COVERED

The three positions covered three different clerical levels. Job A was a kind of Service Clerk/Accounting Clerk mixture, located in the office that processed taxes. Job B was a specialized Word Processing Operator position, initially filled through reclassification of standard secretarial jobs but increasingly filled through outside recruitment. Job C was a highly responsible position acting as primary assistant to the Chief Clerk in an elected official's office.

SCORING APPROACHES COVERED

The three approaches can be summarized as: the "error weighting" approach, the "skills weighting" approach, and the "judgment template" approach. Each approach is described in this paper, together with some ideas for applying it to other selection situations.

THE ERROR WEIGHTING APPROACH

SETTING AND CRITICAL SELECTION FACTORS

The jobs in this case were entry clerical jobs in the tax office. They had heavy turnover and a history of difficulties in selection. For much of the year, people in these jobs performed detail-oriented desk work, adding and checking figures, processing forms, and handling correspondence relating to taxes. For two hectic months, they also staffed crowded information windows and responded to long lines of angry, confused taxpayers, many of whom had unusual names, spoke limited English, or were unfamiliar with the State's tax procedures.

TEST FORMAT

For Job A, the work sample format selected was a set of tasks similar to primary duties of the position. The examination process included an alphabetizing exercise, an exercise involving standard forms and letters, a tax form checking and correction exercise, and an interactive role play exercise.

Each exercise represented an assignment that every new employee would face, with little training, during the first two months of work. They were chosen because they represented assignments where poor performance would lead to an immediate consideration of terminating the employee if they could not handle that assignment.

SCORING AND CUTOFF APPROACH

It was determined that the key issue in gaining buy-off from this office on the examination was to evaluate the seriousness of different kinds of errors that might be made on these assignments. After lengthy debates over possible scoring approaches, we settled on definitions of "major errors" and "minor errors" that could be made on any given exercise. For example, a minor error in the alphabetizing task was defined as any two address cards that were transposed one position from where they should have been. A major error was any transportation more than one position away from its correct place. The supervisor of the filing work indicated that up to three minor errors might be tolerable, given that number of cards to file, but that no major errors were tolerable, since filing errors quickly cumulated and made the files chaotic. Similarly major and minor errors were identified on letters and tax form exercises and also on the role play exercise.

In each case, SME's determined how many minor errors and how many major errors would be tolerated from a new employee. The various exercises were ranked in order of importance, and a final decision was made as to how many major or minor errors, on which exercises, would constitute a screenout on the examination as a whole. Scores were reported to applicants and to the department in terms of these errors, rather than as positive scores. This made it very clear to the department exactly what kind of risks they would face in hiring any individual on the certified list.

RESULTS

The work sample test and the scoring approach were both successful for Job A. Officials indicated that they were seeing a more qualified group of applicants on the certified lists, that they understood exactly what they were getting when they interviewed the applicants, and that the applicants they hired were more skilled at the tasks assigned to new workers, and made far fewer mistakes, than those hired under the previous system. They also found that scores expressed in terms of number of errors, rather than positive points, gave them useful information for selection purpose.

The approach was also helpful in dealing with unsuccessful job applicants, since they could understand exactly what made them fail.

IMPLICATIONS

The error weighting approach was well suited to this situation because the work was so detailed and involved so many repetitive tasks where mistakes could be clearly defined and consequences clearly demonstrated. The weighting of different kinds of errors--in this case in terms of major and minor--allowed Personnel to move away from the preconceptions and speculations engaged in by the selecting officials on the written test, and move towards the standards they actually used to judge employees doing this kind of work. It would be likely to adapt best to use with entry-level kinds of jobs, where applicants might otherwise challenge an assessment of their aptitude, or where selecting officials have been unable to clearly define their selection needs or their reservations about the methods used for selection.

THE SKILLS WEIGHTING APPROACH

SETTING AND CRITICAL SKILL FACTORS

This examination was developed two years after the introduction of a word processing system. This introduction had been gradual and somewhat haphazard. The first machines had been delivered, placed next to the desks of various clerical employees, and after a week or so of training, these employees began doing word processing. As they became more skilled, they were given more work to do. Eventually there were pressures to reclassify the jobs upward, given the additional complexity of the work.

The goal was to define the journey level word processing job based primarily on direct machine skills. Word processing training courses were not yet at a point where completion of training could be used as a standard, and various members of the clerical union believed that there were some employees currently being paid at the journey level who did not possess adequate skills, and others not qualifying at the journey level who did possess them.

A SME committee was formed to assist in planning and developing the tests. This group began by creating a comprehensive outline of all major skill functions on the County's word processing system and taking a survey of current word processing operators to see which functions they knew how to perform and how frequently their jobs called on these skills.

TEST FORMAT

The final format for the examination process was a multiple-choice job knowledge test followed by an on-screen work sample exercise. Both components covered a full range of the word processing skills on the outline; the written test covered knowledge of how to perform various functions, while the performance piece covered skill in applying that knowledge to an actual performance task.

The performance test consisted of one draft letter with handwritten editing notations to be entered into the word processor as a new document, and a three-page report already on the machine that required further editing. Both documents were to be printed after editing.

SCORING AND CUTOFF APPROACH

Most jobs in the jurisdiction were found to require a mixture of basic, intermediate, and advanced level functions. The final approach selected was to score each phase of the test with three subscores, one each for basic, intermediate, and advanced level skills. A pass point was set for each subscore based on the use survey and on pretest results from a sample of experienced journey-level operators identified by the word processing training coordinator as knowledgable at an independent level of functioning on the machine. To pass the test, an individual had to obtain a passing score on each of the three levels. Those who failed one or both of the upper levels could take remedial courses or study their training manuals for those functions and take the test again after a waiting period.

RESULTS

Although there were quite a few problems developing this examination, including the fact that scoring the performance test was time-consuming and difficult, overall the separation of scores by different skills categories or weights was helpful to both applicants and selecting supervisors. Applicants who failed the test received more useful information on what to study than they would have if the test had had a single score.

IMPLICATIONS

This scoring approach can be useful to anyone who is trying to develop a test for skills that are not absolute, but are dispersed unpredictably through either the qualified workforce, or the applicant pool, or both. It is also useful for situations in which the skills base is changing over time.

For any such test, it is critical to have some source of information from a training perspective to assist in defining skill levels appropriately. Part of the rationale behind using a test of this sort was the fact that the knowledge and skill requirements for word processing jobs required fairly extensive training. Very few individuals were able to teach themselves the full range of techniques on the system within a short time period.

The tie-in back to the training materials or programs can also serve to make the test more palatable to applicants, since it can be seen as an aid in diagnosis and career progression, rather than just as a barrier to being hired or promoted.

THE JUDGMENTAL TEMPLATE APPROACH

SETTING AND CRITICAL SKILLS FACTORS

This was a highly political selection situation. The vacant position was in an elected official's office at a high level in the jurisdiction. The previous incumbent was the only person who had ever held the job and was unavailable to interview about the job content. The new selecting supervisor decided to revise the duties, and wanted to give a fair shot at the job to several employees at lower levels in that office, as well as to other employees of the jurisdiction and to outside applicants. Her critical need was for someone who could handle a wide variety of written materials and make appropriate judgments on sensitive or complex issues in her absence.

TEST FORMAT

The examination format finally chosen consisted of a clerical in-basket style exercise. A resource folder was compiled for each candidate including a simplified list of office policies and responsibilities, several schedules and routing lists, and other materials to provide guidance for the exercise. This resource folder was provided to candidates in advance and kept by them for reference during the exercise. Items in the in-basket included correspondence, mail, notes from the supervisor, items to prepare and type such as meeting agendas, replies to correspondence, and phone messages to handle.

SCORING AND CUTOFF APPROACH

The selecting supervisor was interviewed to determine what in her view would be an acceptable approach to handling each item in the in-basket. Her judgments were broken down into three scorable factors: responsive actions, prioritizing and problem analysis, and follow-through. A form was developed to be filled out by candidates as a summary of their decisions in the exercise, and the supervisor prepared a comprehensive summary of all responses that she felt deserved point credit. Points were awarded based on the supervisor's input as to how she would judge the adequacy of responses if the candidate were a new employee. The final score was a total of all points from the template outline. It reflected the degree to which candidates had processed the work and matched the supervisor's judgments. Using the scoring template, each in-basket could be evaluated in about 15 minutes, rather than the hours it could have taken assessor-style.

RESULTS

This test approach was well received by most of the clerical applicants who participated in the examination. They felt it challenged them and gave them a realistic picture of what the work would entail. The selecting supervisor found it helpful in making her final selection. She was given access both to scores and to in-basket folders of each candidate interviewed, and was able to discuss with candidates the reasons for her judgments in the scoring template and the candidates' understanding of the "second-in-command" role of the job.

The candidate selected was not someone the supervisor expected to do well, but performance after selection bore out the high score she received on the test. Other candidates were able to better understand the reasons for their non-selection. Finally, use of the supervisor's template and the objective point-scoring allowed the participation of outside raters for the exam without creating the possibility that their evaluations would widely differ from the supervisor's preferred solutions to the judgment problems in the in-basket.

IMPLICATIONS

Sometimes there are no "right" or "wrong" answers in work samples. As an alternative to using pooled judgments of raters (the assessment center model) for scoring, it may make sense to accept the notion that the supervisor's judgments on handling a problem constitute the most reasonable scoring template. This approach can be used best in situations where this concept will make sense to the applicants, especially for jobs where the coordination between this vacancy and the supervisor's position is extensive.

* * *

The Multiple-Choice In-Basket Exercise as Developed and Used by the New Jersey Department of Civil Service

John C. Kraus, New Jersey Department of Civil Service

Large candidate populations usually preclude a test developer's use of examination modes such as orals, essays and assessment centers. This becomes most acute when testing for middle-to-upper management positions, since those examination methodologies which are usually considered the least efficient are, in fact, often the most preferred. For the State of New Jersey, which maintains a centralized civil service system and is responsible for over 10,000 state, county and municipal titles, this problem is not unusual. Indeed, logistical and fiscal considerations seriously restrict a test developer's options in selecting the appropriate examination methodology and place an over-reliance on the multiple-choice (MC) format. In addition to the candidate population size, the multitude of titles discourages position-specific, multi-part examinations.

Therefore, a methodology was sought which would effectively assess managerial skills and abilities in an efficient manner. The new instrument or procedure would be required to handle large candidate populations and be

more generic in content than traditional position-specific examinations. If an efficient method was devised as a first component, then subsequent multiple parts could be more easily introduced, since the candidate population would be largely reduced.

The primary obstacle was candidate population size. For example, a population of more than 20 candidates is usually considered too large for an oral examination. Similarly, a population of 30 to 60 candidates (depending on the length of the examination) is usually considered too large for scoring an essay examination or case study. The MC format was thought to be redundant as it tended to over-emphasize technical knowledge, an area that someone in a managerial position has probably already been tested on and knows.

The efficiency of the MC examination format for large candidate counts, however, could not be overlooked. Some way was therefore needed to incorporate the MC scoring format into a test product which more closely approximated job behaviors. In determining the actual content of the new methodology, our attention was primarily directed to assessment center exercises. The in-basket exercise quickly became the most attractive choice because it met several criteria: 1) it is the most widely accepted assessment center exercise for measuring the abilities and skills (e.g., planning and organization, judgment, problem analysis) required in the managerial and administrative positions. Indeed, as a work sample and from the perspective of face validity, there is no reason to question the in-basket exercise. 2) the in-basket could easily be made generic in content and used simultaneously for various titles. 3) this exercise appeared to lend itself best to MC answers.

We decided to develop the first MC in-basket for nine different management positions (involving 16 different symbols or facilities) in the social service area. All titles were also scheduled to have a second part examination component, such as an oral or essay, administered at a later date. Only those candidates who passed the MC in-basket would be permitted to take the second-part examination.

As with the traditional in-basket, the MC in-basket consisted of various correspondence, organization charts, background material, "stuffing materials," etc. Although the in-basket was geared to the social service area, it remained sufficiently "generic" in that no technical knowledge of the field was required. Rather, "generic" issues such as promotions, parking problems, disciplinary matters, scheduling conflicts, letters of complaint, budget expenditures, etc. were presented.

In consideration of test administration time constraints and candidate "load," the total number of stimulus items was limited to seventeen. (Subsequent MC-in-baskets have consisted of 15 to 20 items). All the items were numbered and presented in one booklet.

In conjunction with and subsequent to the development of these items or stimuli, MC questions were also being generated. Questions were designed to measure various skills and abilities such as judgment, planning and

organization, and were worded so that they referred back to a particular item or set of items. A total of 32 MC questions were used. (Subsequent MC in-baskets have ranged from 25 to 45 questions). All the MC questions were presented in a booklet separate from the stimuli. It was previously decided that the answers would be determined by pre-testing. That is, three consultants, with excellent management credentials and experience across various agencies within the social service field, were contracted to determine the correct answer choices. This process required each consultant to take the examination and to derive answers independently. In order for a question or item to be retained, all consultants had to agree on the answer.

Pre-testing with the consultants proved invaluable. For example, ambiguity and data inconsistency across the stimuli were identified and corrected. Perhaps more meaningful, however, was the consultants' ability to provide the proper perspective on the stimuli or items presented. That is, some problems or errors which were embedded in the stimuli were found to be too subtle for detection. Other times the consultants claimed that the expected analysis of a particular detail or item was unreasonable in light of the responsibility and level of the position tested (e.g., subordinates, not the manager, would be responsible for examining such detail). The entire consultant pre-test process was quite rigorous and took several weeks to accomplish.

Since the format was a departure from what candidates are led to expect, an explanation was in order. Therefore, six weeks prior to the administration, a letter was sent to all candidates briefly explaining the format, what was being measured, and the process of pre-testing.

More than 400 candidates took the examination across the various titles and symbols. An overall reliability coefficient (K-R) of .70 was achieved for the 31 questions (one question was deleted as a result of its ambiguity). Candidate feedback to this hybrid examination was quite positive, with comments that it was "refreshing, job related and challenging." Negative comments were comparatively few. Only two candidate appeal letters were received which challenged the answers to individual test items. They basically stated that the "answers only reflected the preferred 'style' of the consultants and were not in agreement with management principles." Two instructors of management courses were asked to review these items in response to these appeals and found the appeals to have no merit. However, it was decided that any pre-testing involving future MC in-baskets would also involve an experienced instructor from a management training program.

In addition to the qualitative improvement to our examination product, the MC in-basket has been a resounding success in terms of organizational efficiency. Other intangible factors, such as improved public relations, are also evident. Appointing authorities, in fact, have requested that MC in-baskets be administered for future examination announcements. As a result, five of these exercises have been developed to date. Four are directed at middle-management positions; one for upper-management. Their use within the Department's Division of Examinations has spanned from engineering to accounting managerial titles. Indeed, while MC in-baskets

may require extensive time to develop, their "return" in terms of "milage" or re-use demonstrates that they have been well worth the effort.

An illustration of the in-basket exercises and the MC exam was given.

* * *

ATTRITION: ANALYSIS AND SELECTION-RELATED SOLUTIONS (Paper Session)

Biodata Research Project: The New York State Experience

Glenda K. Corcione, New York State Department of Civil Service

Robert Means, OXICON/McGraw-Hill, Inc.

Introduction/Background

The New York State Department of Civil Service and OXICON/McGraw-Hill, Inc. (a California-based consulting firm) are conducting a two-year research effort to see if biographical data can be used to improve the selection of Mental Hygiene Therapy Aide Trainees (MHTATs).

In New York State, the over 20,000 Mental Hygiene Therapy Aides (MHTAs) constitute the largest number of direct care providers to the mentally and developmentally disabled in over 40 mental health and mental retardation facilities statewide. (Trainees are promoted to Aides upon successful completion of a one-year traineeship).

Mental Hygiene Therapy Aides and Trainees carry out a wide variety of routine and often repetitive tasks connected with the personal care, treatment, and rehabilitation of mentally and developmentally disabled patients. They encourage and guide patients in the development of daily living skills and take care of the patients' personal needs when the patient is unable to do so for him/herself.

New York State's current selection procedure for the Trainee position requires that applicants read, write and speak English, and that they compete in a written examination which tests their understanding of how to care for the mentally ill and disabled. The salary for this entry-level position is $14,000 which is, for most parts of the State, a very attractive entry-level salary to many individuals who have no specialized education or experience.

An attractive entry-level salary, coupled with no specialized education or experience requirements may cause one to wonder why New York State is concerned with improving the selecting of MHTATs.

It appears that although MHTATs are aware that promotion to journey-level status is dependent on successful completion of a rigid training program, they are unaware or unprepared for the distasteful and frequently, stressful aspects of the job including changing diapers on adults, warding off abusive behavior and spending months teaching adults basic daily living skills. This mismatch of people to jobs has led to significant performance and tenure problems in the first year after hire. This, in turn, translates to high costs in recruitment, training, and counseling, not to mention the decreased quality of care to patients, and overwhelming cost to taxpayers.

In addition, morale among current employees is low. While Trainees are in classroom training, an unreasonable burden is placed on current staff who are forced to care for more patients than normally planned for and who are forced to work overtime when coverage on the next shift is insufficient. This develops into a vicious cycle, causing absenteeism due to illness and fatigue, which causes more overtime and possibly a lower level of performance for those remaining Aides and Trainees.

New York State is attempting to address the problems of poor performance and high turnover for these positions by researching an alternative selection mechanism which has the potential of "matching" applicants to the MHTA position. During the first year of the research study, a biographical questionnaire was developed which appears to predict, without adverse impact on protected class members, the high performance and long tenure probability of candidates for the MHTAT position. The second year of the study, currently underway, will provide New York State with enough additional information to determine whether the results from the first year can be generalized to future MHTA applicant populations.

Biodata - Definition and Use

Biodata is a multi-purpose process based on the premise that past behavior is predictive of future behavior. It captures an individual's motivation attributes, measuring affective, not cognitive, needs. It addresses the question, what drives a person?

Operationally, biodata matches an applicant's background, experiences, and preferences against that of a composite profile of successful incumbents to yield an objective measurement of an applicant's fit for a job.

Biodata has been used for employee selection in the private sector and public jurisdictions and has been shown to be predictive of performance and tenure, without adverse impact. It has been used in the private sector for over 60 years for such titles as bank tellers, engineers, sales representatives, and managers. More recently, biodata has come into use in public jurisdictions for such titles as eligibility workers, clerks, and correction officer trainees.

Overview - Component Parts of Research

In researching biodata for any title, data must be gathered from three major sources: a biographical questionnaire that incumbents and applicants respond to; a performance evaluation on the incumbents who provided those responses to the questionnaire; and turnover data information. From those components can be derived a performance profile and a turnover profile can be derived.

For the performance profile, the responses of the top performers will be compared to the responses of low performers to find out how they are different. For the turnover profile, the responses of long tenure incumbents will be compared to the responses of short tenure incumbents to find out how they are different.

After the profiles have been developed, the questionnaire may be administered to applicants, and their responses would be compared to the aggregate profile of successful incumbents. The closer an applicant's responses match those of the successful incumbent profile, the higher the score will be.

The authors discussed how the biographical questionnaire and performance evaluation form was developed during the first year of the project. The last step was to match the incumbent response form with that incumbent's performance evaluation form (for purposes of the performance profile). After removing all the problem cases, the matching process began. As one might expect, not all the biodata response sheets had matching performance evaluation forms and not all the performance evaluation forms had matching response sheets. Consequently, of the 6,576 response sheets and 7,360 performance evaluation forms submitted, only 3,693 actually matched, an "n" sizable by normal standards, but constituting a relatively small percentage of the population.

General Survey Findings

When comparing applicant and incumbent respondents, the research revealed that applicants were considerably younger and had more directly relevant prior experience, but had only slightly more formal education. In both applicant and incumbent groups, minorities were comparably represented and males and females similarly represented.

When comparing job performance dimension weights and overall performance evaluation scores between mental health and mental retardation facilities, the research revealed that the average importance attached to the performance dimensions and the distribution of performance evaluation scores was quite similar.

When comparing performance evaluations with other factors, age more than tenure is correlated with performance evaluations, educational level is <u>not</u> correlated with performance evaluations, and females and whites receive somewhat higher evaluations than males and minorities.

Second Year

The second year of the project currently underway, is a simple predictive follow-up study. The volume of individuals involved is cut in half, thereby reducing the administrative complications.

Conclusion

New York State believes biodata offers significant potential for improving the screening and selection of Mental Hygiene Therapy Aide Trainees. The instrument developed during the first year appears to predict performance and tenure without adverse impact on protected class members. Although the samples are a relatively small percentage of the population, the samples are, as mentioned earlier sizeable by normal standards and the statistical results are consistent across samples.

The second year of the study should provide New York State with enough additional information to determine whether the results from the first year can be generalized to future applicant populations. New York State Department of Civil Service will then be in a position to determine whether biodata will be used in selecting future Mental Hygiene Therapy Aide Trainees.

* * *

Police Dispatcher: An Analysis of Attrition

George Rost, City of Los Angeles Personnel Department, Los Angeles, CA

Until 1982 the Los Angeles Police Department employed Radio Telephone Operators who took written instructions from officers and then dispatched patrol cars. At that time the decision was made to install a new computer dispatch system and to civilianize the communications operation. The City established a new class of Police Service Representative (PSR) to do both functions - take calls from the public and dispatch patrol cars. Also the 911 emergency system would be made operational. The new system using 911 did not go into operation until 1984 after significant hardware problems.

As a result of this changeover, 193 PSR's were hired in 1984. One hundred five of them left during training. In 1985 the Police Department established a committee to study the attrition problem and invited us to join it. The Personnel Department then decided to do a study to analyze attrition.

The study included:

1) Survey of other jurisdictions
2) Analysis of test results for the 193 PSR's hired
3) Interviews with PSR's
4) Job analysis and typing requirement

Richard Mancuso and Sandi Peelen of our staff did most of the work on the study and their contributions are gratefully acknowledged. Rich prepared most of the staff reports that formed the basis of this report.

I. Survey of Other Jurisdictions

Method

We developed a questionnaire designed to gain a picture of police dispatching and attrition in jurisdictions using civilian dispatch personnel. The survey consisted of 31 questions and covered a variety of dispatch related topics including demographics, selection procedure, recruitment practices, turnover rates and training. Results were gathered from 14 jurisdictions.

Results

The four primary pre-employment testing procedures used by surveyed agencies are written tests, oral interviews, typing tests and, in 57% of reported cases, simulation or performance tests.

While the basic test types exist, there appears to be minimal consensus on testing specifics. Most frequently tested abilities lay within the memory, verbal, and following directions domains. Surprisingly few agencies indicated that they tested "decision making" or prioritizing" abilities directly.

Reported training attrition rates appeared, for whatever reason, to be clustered into three groups: A low level (15% and below), moderate level (20-30%), and a high level (40% or greater).

No single, identifiable factor emerged from the survey data to explain why some agencies experience low training attrition rates while some had high rates.

While most agencies used a typing test, there were significant differences in speed requirements. Interestingly, of the two agencies using the highest speed levels (45 WPM), one reported a 50% training attrition rate while the other reported 20%. Clearly, typing ability does not guarantee training success although most agency personnel believe it to be a important job related factor.

II. Analysis of Test Results for the PSR's Hired in 1984

Method

Data for the study were gathered for all persons who entered PSR training classes between January 30, 1984 and January 20, 1985. The total of 233 individuals was distributed among 6 classes.

Data on each individual's race, sex, written, oral and final selection scores were collected from Personnel Department records. A determination of whether the individual took the entry exam on an Open or Promotional basis was also made. Attrition statistics as well as the total number of months attritees remained in the PSR program were gathered from PSR training records. In addition, personnel records folders for each of the 233 trainees were individually reviewed to determine which of a potential eight occupational categories an individual had held prior to PSR employment. For clerical personnel, a determination was also made as to whether pre-employment experience was at a supervisory or non-supervisory level.

All data were computer analyzed using the various facilities and program routines offered through SAS (Statistical Analysis System).

Results

Sex, Race, Ethnic Distribution and Examination Status

The 233 class members included in the study contained 92% females and 8% males. A majority were minority group members. A majority of the study population (61%) had taken the PSR examination on a promotional basis while 39% came from outside City employment.

Attrition, Frequency, and Tenure

Those individuals leaving the PSR program did so at varying points during their training and probation. The largest single percentage of the 109 attritees terminated their employment after 3 months on the job. The 5 and 7 month points resulted in the second highest attrition rates. Fully 1/3 of all attritees had terminated within the first 3 months of employment while more than 2/3 of the trainees left prior to completing six months of training. Less than 1/5 of attrition occurred after 7 months on the job.

Examination Status and Attrition

Two significant differences between "Open" (non-City employees) and "Promotional" (City employees) candidates emerged. The groups showed sizable differences in their rates of attrition and in their average tenure prior to attrition. The promotional candidate termination rate for the classes examined was 65.1% while the "open" candidate rate was 40.3%. The rate difference was statistically significant. (Chi Sq = 15.55 p .0001). Average tenure for attritees also differed

significantly. Excluding the January class, "open" candidates remained an average of 6.2 months while "promotionals" stayed an average of only 4.8 months. ($T = 2.40$ $p = .017$).

The most likely explanation for open-promotional differences is the ease of acquiring new employment between City and non-City employees. Clearly, when confronted with any of the difficulties leading to attrition, the ease of obtaining a new job affects both the decision to leave and the speed with which that decision is made. The significant difference in months on the job between "opens" and "promotionals" (6.2 months vs. 4.8 months) lends further support to this conclusion.

Pre-Employment Experience Category and Level of Responsibility

All trainee pre-employment applications were examined, and the applicant placed into one of eight job categories based on the nature of their most recent job prior to taking the PSR test. For clerical positions a further determination was made (based on the candidate's application) as to whether the job was "supervisory" or "non-supervisory" in nature. Attrition rates were then calculated for job type and level of supervision. Rate differences were then compared to determine if prior employment had any effect on attrition.

No statistically significant relationship was found between recent pre-PSR occupation and attrition (Chi Sq = 7:67 p .05). When corrected for small cell size, even the fairly substantial difference in attrition rates between former Clericals (63%) and former Police Dispatchers (33%) failed to meet the parameters established to assure confidence in results (Chi Sq with Yates Correction = -3.68, p .05).

Absence of a Test-Attrition Relationship

A major focus of this study centered on an exploration of the relationship between attrition and pre-employment test scores (both written and oral). Although, definitional imprecision and statistical problems (range restriction), may have acted together to make any test-attrition relationship difficult to detect, the fact remains that no significant relationship was found.

III. Interviews with PSR's

Method

Interviewees were selected randomly from several categories. Tenure ranged from 30 years (RTO-PSR) to 6 months (trainee). Interviews lasted an average of 1 1/2 hours. Two similar lists of open ended questions were developed for the interviews. Instructors and Senior PSRs were asked one set while current and former PSRs in a non-supervisor or non-instructing capacity were asked another. Each set of questions contained a core of identical questions with additional items focused on issues of particular relevance to each group.

Overview

Several major reasons for PSR attrition emerged from PSR interviews. Among incumbents (combined Instructor/Supervisor and PSRs) three reasons were mentioned with equal frequency: unrealistic pre-employment job expectations, the rigid and generally inflexible work schedule and poor instructor student relationships. Each was mentioned by 43% of incumbent interviewees. When former employee opinions are added to those of incumbents, poor instructor-student relationships emerged as the most frequently mentioned reason for PSR attrition (54%, 14 of 26).

Other reasons for attrition cited by incumbents were: adjustment to shift work (28%), job induced stress (14%), inadequate trainee ability and skill levels (14%), and poor pre-employment testing (10%). Differences in the frequency with which each attrition contributor was mentioned emerged between Senior/Instructors, PSRs and former employees. Supervisors and Instructors placed much greater emphasis on the problems presented by scheduling, inflexibility and rigidity (63%) and unrealistic pre-employment expectations (55%) than they did on instructor-student problems (36%). On the other hand, one of every two non-instructor/supervisor incumbents mentioned instructor-student relationships as a major cause of attrition while 100% of the former PSRs cited instructor-student relations as a major cause of attrition.

Training Environment

When all interview groups are considered, the most frequently mentioned cause for PSR attrition was poor instructional atmosphere. In particular, a strikingly negative relationship between students and a significant percentage of instructors was cited by 50% of non-instructor/supervisor, 100% of former employees, and 36% of Instructor-Supervisors. A substantial number of interviewees described incidents in which trainees were treated in an abusive, derogatory or humiliating manner by floor instructors. A surprisingly large number of interviewees characterized the learning atmosphere as one of constant emphasis on mistakes and errors, with little positive reinforcement for achievement.

Some interviewees attributed what they felt was a "tolerance" for poor instructors to a continuing instructor shortage. Others felt that eliminating the 10-30% of "bad" instructors would create severe shortages in all areas of PSR operation. Some felt the instructor selection system was ineffective and focused on job knowledge exclusively at the expense of teaching ability. The lack of a standardized curriculum for floor trainers and significant differences in teaching technics were also cited.

Skills, Abilities and Traits Viewed as Necessary for PSRs

In the areas of equipment operation, decision making, and communication skills, all groups felt that equipment operation was the least important of the three job components, while decision making and communication skills were essentially of equal importance in PSR job performance. In round numbers, most interviewees felt that approximately one-quarter of the job involved familiarity with and operation of equipment while 3/4 of the job entailed making decisions, gathering and transmitting information. Non-instructor PSRs tended to place slightly more emphasis on equipment and emphasize communication skills over decision making as did former employees, while Instructor-Supervisors placed less emphasis on equipment operation and emphasized decision making more than the other groups.

Stress

In addition to the other attrition contributors mentioned by interviewees, the effect of job related stress was spontaneously mentioned by nearly half (40%) of the non-instructor/supervisors, though not by the other groups. However, when specifically asked about the role stress played in the job nearly all respondents commented on the stressful nature of their jobs. When directly questioned directly on the causes of stress some interviewees mentioned civilian versus sworn problems. Others mentioned constantly changing procedures, rules, etc; the lack of a forum for expressing frustrations to management; "second guessing" of decisions and the fear of making mistakes. For Instructors, the constant unending flow of students with accompanying lack of relief from teaching was a clear contributor to Instructor stress and "burnout." Interviewees indicated that it was the "work added" stress which was problematic, rather than the nature of the job itself.

Interview Conclucions and Recommendations

The consistent citation of instructor-student problems by all interview groups (including instructors) clearly identified an area that needed substantive corrective action (instructional environment).

More than half (55%) of Instructor-Supervisors and nearly 1/3 (30%) of other incumbent interviewees pointed to unrealistic, inaccurate or unclear new hire job perceptions as significant attrition causes. While not as frequently mentioned as learning environment and unrealistic expectations, inadequate trainee skill levels and lack of adequate pre-employment testing were cited by 27% and 18% of Instructor-Supervisors and non-supervisors respectively as major attrition contributors.

IV. Job Analysis and Typing Requirement

Method - Job Analysis

Our staff observed the Investigating Report Operator positions, the Emergency Board Operator position, the Radio Telephone Operator positions as well as the classroom training of new Police Service Representatives. We then held a series of task identification, element identification and rating meetings with Police Service Representative, instructors and supervisors.

Results

Seven factors were identified as critical to be examined in a written test:

1. Ability to organize data and make decisions/put information in priority order
2. Ability to follow rules, steps or procedures and apply them to specific instances
3. Ability to follow oral directions and record numeric information
4. Ability to communicate, listen, retrieve information using correct vocabulary and grammar
5. Ability to match alpha and numeric data
6. Reading comprehension
7. Memory

Method - Typing

The police department maintained that it is necessary to be able to type faster than 30 wpm when entering data into the console while taking emergency calls. We had, however, observed some PSRs with poor typing skills who had been on the job for a long time with apparent success.

We prepared a typing test of standard report material and a cassette tape of special material, mostly names, descriptions and vehicle license numbers, played it on the console, and they typed directly into the computer.

Results

The results for the written copy were somewhat better than we expected, a mean net of 43.8 wpm on the IBM Selectric and 49.9 on the console. We had considerable difficulty scoring the taped material because all of the PSRs used abbreviations in typing descriptions. We finally scored it as a percentage of correct key strokes and did not count abbreviations as errors. The results of the oral typing test correlated positively with the other typing tests. We then felt that we had established the point that if we tested candidates on our IBM Selectric typewriters they should be able to adapt to typing from verbal instructions. We set the cutoff at 32 net wpm which was one standard deviation below the mean.

Overall Recommendation and Results

Personnel Department Responsibilities

1. Prepare a realistic job preview check list. Checklist sent to all candidates for 1986 examination with test notice. About 40% lapse rate which was slightly higher than before. Feedback indicated that it served as a good means of informing candidates.

2. Prepare a written test based on the job analysis. New test administered in January 1986. Good candidate acceptance and excellent police department acceptance.

3. Administer typing test with new 32 net wpm requirement. Thirty-eight of 44 passed the March 1986 testing. Higher pass rate than before.

Police Department

1. Reduce class size from 40 to 20
   - January 1986 class - 20

2. Identify, correct or remove abusive instructors
   - They identified some and shifted them to other assignments

3. Schedule instructors for regular non-instruction periods to avoid burnout
   - Some instructors on sabbatical due to smaller class size

4. Purchase and use the same equipment in training as is used on the job
   - Not implemented yet

5. Hire a training consultant to review: instructor selection criteria; letter training for instructors; curriculum design; and standardize floor instruction
   - being reviewed

6. Maintain instructor trainee assignment throughout training
   - being implemented

7. Provide trainees with fixed time and rotation schedules minimum of a month in advance
   - being reviewed

Final Comments

The Police Department cooperated with our study throughout because they recognized that with a staff of almost 400 Police Service Representatives and the responsibility for an emergency communication function it was essential to have an effective selection and training system. I believe that we are well on the way toward that goal.

PSYCHOMETRIC ISSUES AND TECHNIQUES (Paper Session)

Using "Lemon" Job Analysis Tasks in Examination Validation: A Technique

Catherine S. Cline, New York City Department of Personnel

Job analysis questionnaires may be administered to employment applicants as part of "training and experience" selection examination, as a screening mechanism, or as part of general construct validation of an examination. In all cases, the hypothesis inferred or explored is that previous performance of tasks relates to future performance in the position. A general problem with these questionnaires, as with all self-report instruments, is that applicants may inaccurately report previous experience.

In the present study, a questionnaire of task statements was administered to candidates for a managerial position within a civil service agency. Candidates indicated whether they had a) not previously performed each task; b) performed it only under supervision; c) independently performed it, or d) supervised it. Previous data indicated all tasks, except two, were critical to the position. The two "lemon" tasks were tasks job experts agreed are not performed either in the managerial position, or in its feeder titles. Questionnaires were voluntarily completed by 46 candidates immediately prior to administration of an in-basket and essay exam, constructed to assess the questionnaire task dimensions.

Table 1 presents inter-rater reliability and mean rater scores for seven out of 12 in-basket tasks administered to the candidates. Inter-rater reliability for these tasks was in the 80's and 90's, showing a sufficiently detailed scoring protocol, the remaining five tasks were not used as criteria in this study because scoring had not been completed or reliability was low.

Table 2 shows the number of persons endorsing each or either of the two lemon items included on the 29 statement task questionnaire. The lemon items asked candidates whether they reviewed FHAC (a neologic acronym) regulations for impact on existing policy, and if they prepared unit budgets. It is noteworthy that approximately half of the sample endorsed one or the other of the "lemon" items, even though they completed the questionnaire on a voluntary and confidential basis.

Tables 3 and 4 present the in-basket performance of the endorsers and non-endorsers in raw and standard score forms respectively. Persons who endorsed at least one lemon item scored lower on all in-basket tasks than persons who did not endorse lemon items. On five out of the seven tasks the difference between the two groups was significant.

Table 5 shows the mean ratings on the questionnaire items for endorsers and non-endorsers of lemon items. Endorsers of lemon items had a mean self-report rating of 3.36 on the genuine tasks contained in the questionnaire; i.e. they reported that they had either performed the tasks independently or

had supervised them. Non-endorsers had a lower mean, reporting they had performed most tasks either under supervision or independently. Finally, Table 5 shows that questionnaire results, as might be expected, were much less reliable for endorsers that for non-endorsers of lemon items.

Results indicated that endorsers of lemon items (almost half of the sample) performed more poorly on in-basket tasks than non-endorsers of lemon items. In contrast to their actual performance, endorsers also rated themselves more highly on a T&E - like questionnaire, indicating a serious problem with this type of assessment.

TABLE 1

INTER-RATER RELIABILITY OF SELECTED IN-BASKET TASKS

TASK	RATER MEAN 1	RATER MEAN 2	INTER-RATER RELIABILITY	NUMBER OF CANDIDATES
1.	5.54	5.97	.868	144
2.	5.41	5.67	.913	144
3.	5.85	6.46	.830	139
4.	8.84	9.48	.885	143
5.	2.79	2.81	.907	138
6.	11.13	11.46	.954	134
7.	4.62	4.61	.923	140

TABLE 2

ENDORSERS AND NON-ENDORSERS OF LEMON SURVEY ITEMS

LEMON ITEM	NON-ENDORSER	ENDORSER	PERCENT ENDORSERS
FHAC	29	12	29.3
BUDGET	29	13	31.0
EITHER ITEM	24	21	46.7

TABLE 3

IN-BASKET PERFORMANCE OF ENDORSERS AND NON-ENDORSERS OF LEMON ITEMS

TASK	NON-ENDORSERS (N=29)	ENDORSERS (N=21)
1.	6.09	5.05
2.	5.56	5.23
3.	6.64	5.58
4.	9.22	8.62
5.	3.13	3.00
6.	12.08	9.76
7.	5.38	4.15
Rating Average	6.87	5.95

TABLE 4

COMPARISON OF ENDORSER AND NON-ENDORSER IN-BASKET

STANDARDIZED AVERAGE RATINGS (1)

TASK	NON-ENDORSER	ENDORSER	DIFF.	F
1.	+.156	-.442	.598	p less than .01
2.	+.104	-.125	.229	(ns)
3.	+.189	-.215	.404	p less than .05
4.	+.242	-.074	.316	p less than .05
5.	+.236	+.143	.093	(ns)
6.	+.196	-.388	.584	p less than .01
7.	+.395	-.246	.641	p less than .01

(1) To equate across different in-basket tasks, ratings were converted to z-scores using the overall group mean and s.d. within each task. Thus ratings reported here are in s.d. units

TABLE 5

MEAN RATINGS AND RELIABILITIES OF SELF-REPORT RATINGS

OF ENDORSERS AND ENDORSERS

SELF-REPORT QUESTIONNAIRE	NON-ENDORSERS (N=29)	ENDORSERS (N=21)
MEAN	2.87	3.36
SD	1.52	.67
RELIABILITY (Coefficient Alpha)	.868	.381

* * *

Using and Evaluating Ranked Assessments:

The Practical and Statistical Significance of Rank Order Correlations

Andrew S. Imada, University of Southern California,
Institute of Safety and Systems Management

Introduction

Often we use rank order data to predict some future event. By correlating this rank order with a criterion, we can estimate the predictive efficiency of the ranked data or predictor. Peer rankings have been effective predictions of future performance and are thought to be better than psychometric procedures (Kane & Lawler, 1978, Korman, 1968; Lindzey & Byrne, 1969; Miner, 1968; Korman, 1968). Kane & Lawler (1978) distinguished between peer nominations, peer ratings, and peer rankings with the ranking technique being most discriminating and more reliable than peer ratings (Love, 1980). Kane & Lawler found, while Lewin & Zwany (1976) reported a median validity coefficient of .41 for 15 validity studies. However, of the three peer assessment techniques, least is known about the psychometric properties of the ranking technique. This raises serious validity questions. This paper demonstrates how ranked data can produce spurious correlations: Hypothetical, but plausible rankings are presented and explained and methods for interpreting these results are offered.

Situations Likely to Produce Spurious Correlations

Typically, the rankings method requires that each judge rank every group member on one or more dimensions. These rankings are then correlated with some criterion measure using either Spearman's rho or Kendall's tau (See Winkler & Hays, 1975). However, correlation estimates assume linear and homoscedastic relationships, but this is not always the case. There are reasons to suspect that these assumptions are violated, thus accounting for systematic method variance. Three simplified ranking situations are presented to illustrate these points.

Situation 1

A judge is asked to rank order 12 peers on the criterion--of leadership. He can accurately identify the best and worst leaders but is unable to rank the remaining 10 peers. That is, the correlation for the first and twelfth positions is 1.0 and that of the second through eleventh positions is .006. This hypothetical ranking is presented under Judge 1 in Table 1.

TABLE 1

Ranking of Criterion Variable and Three Hypothetical Rating Situations

Criterion	Judge 1	Judge 2	Judge 3
1	1	1	2
2	4	2	5
3	11	3	6
4	2	4	1
5	8	5	3
6	10	12	4
7	5	6	7
8	7	7	8
9	3	11	9
10	9	9	10
11	6	8	11
12	12	10	12

Spearman's formula for rank order correlations estimates that the correlation between the judge's rankings and the actual rankings on the criterion is .41!(1)

In Cronbach's (1955) terms, Judge 1 has effectively utilized differential accuracy when assessing the extremes but failed to do so when ranking the middle positions. A study by Lewin, Dubno & Akula (1971) indicated that the first and last rankings were more accurate than the middle rankings. These results are presented in Table 2.

Situation 2

This situation involves a judge who is able to correctly rank order five peers who are the highest on some measured dimension; but is unable to rank order the remaining 7 peers. The correlation for the 6th through 12th positions is .000 while that of positions one to five is 1.0 (See Judge 2 on Table 1). The rank order correlation for all 12 positions is an impressive .80.

Situation 3

The third ranking situation is similar to the second except that this judge can correctly identify the bottom half of the distribution, but not the top half. The correlation for rankings 7 through 12 is 1.0, but the correlation for positions 1 through 6 is virtually zero. The correlation coefficient for all ranks is impressive—.87 (See Judge 3 in Table 1).

While our hypothetical judges may represent extreme examples, they demonstrate simply, but effectively, the consequences of violating the statistical assumptions underlying the correlation coefficient. We thereby suggest that the "significance" of the correlation coefficient needs more

than a statistical criterion and depends on the intended use of the prediction. For example, if the goal of the ranking is to select the five highest performers, then the 0.8 correlation for Judge 2 is actually an underestimate; it should be 1.0. By contrast the .42 in the first rating situation is a gross overestimate of the first judge's predictive powers. The problem arises when we rely solely on a correlation coefficient to assess the predictive accuracy of the judge (See John Tukey's, 1969 summary of this problem).

Explanations for these spurious rank order effects can be found in several different areas of psychology.

Distance Effects

In his paper on stereotype development, Campbell (1967) posited that the greatest contrast between people being rated will provide the strongest stimuli. In rating one's peers, the best and the worst performers will be most heavily contrasted, and, consequently, ranked most accurately. Work in experimental psychology on the distance effect in comparative judgments provide similar predictions (See Potts, Banks, Kosslyn, Moyer, Riley & Smith, 1978).

The distance effect holds that when presented with different stimuli and asked to make judgments about these stimuli, reaction time increases systematically with the similarity of the stimuli being compared. Conversely, the farther apart two stimuli, the shorter the reaction time. This effect has been noted consistently with chromatic stimuli (Hermon, 1906), visual size comparison (Curtis, Paulos & Rule, 1973), visual numerosity (Lemmon, 1927; Buckley & Gillman, 1974), tonal comparisons (Hermon, 1906), and kinesthetic tasks (Crossman, 1955). Distance effects are observed even when there are no actual physical stimuli and subjects are required to make comparisons of two stimuli stored in memory. Potts et. al. note that "reaction time generally increases (in a Weber-Fechner fashion) as the ratio of physical differences decreases, even when the absolute difference is held constant, and that a small effect of absolute difference is sometimes observed, even when the ratio is constant." (p. 245).

It appears then that people have more difficulty making comparative judgments when the stimuli are similar than when stimuli are dissimilar. The very nature of the distribution can contribute to the distance effect in explaining the ranking of Judge 1 when one compares persons at various standard deviation points within the distribution.

TABLE 2

PROPORTIONS OF POOL OBSERVED AGREEMENT ON QUESTIONNAIRE ITEM FOR DIFFERENTIALLY RANKED INTERACTING GROUP MEMBERS*

Item	Question 1 With whom can you work best?	Question 2 Who contributed most to achieving the goals of the team?	Question 3 Who contributed most to the analysis and solving of day-to-day problems?
Interacting group Ss ranked first	.65	.69	.71
Interacting Ss ranked either second, third or fourth	.31	.31	.34
Interacting Ss ranked last	.85	.87	.89
X2	15.0	11.3	9.3

Note——N=97 for four observer groups. Interacting group, N=14, df=13, ns

*Reprinted with permission from the author.

Lewin et. al. essentially treated the middle rankings as error variance.

Guion (1983) has warned against our sole reliance on correlation coefficients:

> "People place too much faith in validity coefficients; there seems to be a natural tendency to overlook the possibility that nice validity coefficients might be found because both the instrument being validated and the criterion share common contamination...Validity coefficients are, of course, important evidence in making judgments of validity, but one should never confuse a validity coefficient with validity, and one should never base a judgment of validity on a validity coefficient alone." (pp. 6-7).

Psychological Explanations for Spurious Rank Order Correlations

The effects created by Judges 2 and 3 can also be explained by the distance effect and the distribution of people. If the distributions are heavily skewed in one direction, comparisons would be easier at this end than the other end where most of the people are located over a very narrow range of values. Thus, it appears that both the nature of the distribution and the distance effect can explain these spurious rank order correlations.

The Availability Heuristic

Tversky and Kahneman's (1973) concept of availability can be used to explain why extremes are ranked more accurately. In this concept, an event is judged likely or frequent if it is easy to recall relevant instances. However, since availability is affected by subtle factors unrelated to likelihood, reliance on it could result in systematic overestimation for familiar, recent, emotionally salient, or otherwise memorable events. This availability heuristic predicts that certain behaviors would therefore carry undue weighting causing individuals at ends of the distribution to be perceived even more distant from the norm.

Recommendations

If the concerns raised in this paper are real, there are a number of changes that should be made in measuring, interpreting, and using ranked data.

Measurement

While these ideas may not be new, they have not yet been incorporated as an overall strategy for solving the problems addressed in this paper. The first strategy capitalizes on the distance effect. As Campbell (1967) points out, if people can identify the extremes more easily, the task can be structured to get the judge to systematically rate the remaining persons, (i.e., those left after identifying best and worst persons), on the basis of the greatest contrast between them.

The second strategy is based on two well established findings. The first is that judges can recognize only a limited number of entities simultaneously. Miller's (1956) 7 plus or minus 2 seems to be the accepted limit. The second model is March and Simon's (1958) notion of bounded rationality which proposes that people can only do one or a few things at a time, and, that people attend to only a small part of the information presented by the environment and recorded in memory. People deal with information by "chunking" it and use sequential, rather than simultaneous, decision processes. The ranking task can thus be divided into subranking tasks, the subgroups of which can then be assembled sequentially to form an overall rank order.

Interpretation

There has been an over reliance on classical hypothesis testing and significance levels. The fact that a mean difference or correlation is statistically significant does not ensure that it is practically significant.

We need to go beyond correlations. As Cronbach and Gleser (1965) point out, tests and measurements are useful to the extent that they help us make decisions. Psychometric criteria are indeed important in decision theory, but it is the outcomes--to organizations and individuals--that are of prime importance. We need to look at consequences.

Usage

When we first correlated our rank ordering of peers to our criterion variable, we assumed that this correlation expressed the ranking's validity or common variance between our peer rankings and the criterion (Guion, 1983). However, our correlation coefficient is only one measure to express this relationship. Perhaps we need to look at other estimates.

The appropriate parameters will depend on the goals and nature of the decisions to be made. If the data are to be used to select, eliminate, or single out individuals, then we need to look at relevant ranking positions.

An often overlooked parameter is the standard error of the criterion measure. When correlations are low, it is often assumed that the predictor is not accurately predicting the criterion variable. Little consideration is given to the possibility that the criterion may be contaminated or that we may need multiple measures of the criterion variable.

In summary, it seems that the rank order correlations generated in the peer ranking literature may be due to violations of the assumptions underlying the correlation coefficient. To more accurately assess the predictive accuracy of rank order correlations, we may have to go beyond statistical significance by being more concerned with practical significance and the impact of our statistical effects on people and organizations.

References

Buckley, P.B. & Gillman, C.B. (1974). Comparisons of digits and dot patterns. Journal of Experimental Psychology, 103, 1131-1136.

Cascio, W.F. & Silbey, V. (1979). Utility of assessment centers as a selection device. Journal of Applied Psychology, 64, 107-118.

Campbell, D.T. (1967). Stereotypes and the perception of group differences. American Psychologist, 22, 817-829.

Cook, M. & Smith, J.M. (1974). Group ranking techniques in the study of the accuracy of interpersonal perception. British Journal of Psychology, 65, 427-435.

Cowles, M. & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553-558.

Cronbach, L.J. (1955). Processes affecting scores on 'understanding others' and 'assumed similarily'. Psychological Bulletin, 52, 177-193.

Cronbach, L.J. & Gleser, G.C. (1965). Psychological tests and personnel decisions (2nd ed.). Urbana: University of Illinois Press.
Crossman, E.R.F.W. (1955). The measurement of discriminability. Quarterly Journal of Experimental Psychology, 7, 176-195.
Curtis, D.W., Paulos, M.A. & Rule, S.J. (1973). Relation between disjunctive reaction time and stimulus difference. Journal of Experimental Psychology, 99, 167-173.
De Soto, C.B., London, M. & Handel, S. (1965). Social reasoning and spatial paralogic. Journal of Personality and Social Psychology, 2, 513-521.
Guion, R.M. (1983, August). The ambiguity of validity: The growth of my discontent. An address to the Division of Evaluation and Measurement at the meeting of American Psychological Association, Anaheim, CA.
Hermon, V.A.C. (1906). The time of perception as a measure of differences in sensations. Archives of Philosophy, Psychology, and Scientific Methods, 8, 1-75.
Hyde, J.S. (1981). How large are cognitive gender differences? A meta-analysis using omega square and d. American Psychologist, 36, 892-901.
Kane, J.S. & Lawler, E.E. III, (1978). Methods of peer assessment. Psychological Bulletin, 85, 555-586.
Korman, A.K. (1968). The prediction of managerial performance: A review. Personnel Psychologist, 21, 295-322.
Lawshe, C.H., Bolda, R.A. (1958). Expectancy charts: Their use and empirical development. Personnel Psychology, 11, 353-365.
Lemmon, V.W. (1927). The relation of reaction time to measures of intelligence, memory and learning. Archives of Psychology, N.Y. 15, No. 92.
Lewin, A.Y., Dubno, P. & Akula, W.G. (1971). Face-to-face interaction in the peer nomination process. Journal of Applied Psychology, 55, 495-497.
Lindzey, G. & Byrne, D. (1969). Measurement of social choice and inter-personal attractiveness. In G. Lindzey and E. Aronson (Eds.), Handbook of Social Psychology, (Vol. 2), Reading, MA: Addison-Wesley.
Love, K.G. (1981). Comparisons of peer assessment methods: Reliability, validity, friendship bias and user reaction. Journal of Applied Psychology, 66, 451-457.
March, J.G. & Simon, H.A. (1958). Organizations, New York: Wiley.

Footnotes

**Requests for reprints should be sent to Andrew S. Imada; Human Factors Department; Institute of Safety & Systems Management; University of Southern California; University Park; Los Angeles, CA 90089-0021.

The author would like to thank Michael Oakes for his stimulating discussions in the early development of this paper and to Dixie Imada for her support and comments during revisions of this paper.

(1) In light of the above example, the median validity coefficient of .41 in the 15 validity studies reviewed by Lewin and Zwany becomes less impressive.

Maxwell, S.E., Camp, C.J. & Arvey, R.D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.
Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits to our capacity for processing information. Psychological Review, 63, 81-97.
Miner, J.B. (1968). The early identification of managerial talent. The Personnel and Guidance Journal, 46, 586-591.
Potts, G.R., Banks, W.P., Kosslyn, S.M., Moyer, R.S., Riley, C.A., & Smith, K.H. (1978). Encoding and retrieval in comparative judgments. In N.J. Castellas & F. Restle (Eds.) Cognitive theory, (Vol. 3), New York: Earlbaum.
Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24, 83-91.
Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232.
Winkler, R.L. & Hays, W.L. (1975). Statistics: Probability, inference and decision (2nd ed.), New York: Holt, Rinehart & Winston.

* * *

IPMAAC INVITED SPEAKER

Touring Performance Appraisal in a Time Capsule (1)

Gary B. Brumback, U.S. Department of Health & Human Services
Washington, D.C.

Wearing my new T-shirt with CAPTAIN APPRAISAL printed on it, I am going to pilot you in my time capsule for a one-hour tour of performance appraisal (PA).

Here is our itinerary. At hypersonic seed we will travel through Past, taking snapshots along the way as we move quickly to Now where we will visit the Land of Myth and Folly, Battlefield and the Land of MBR. We will stay overnight at Ishu Inn, get up and streak to Future for a quick look and then land safely back home. We will maintain a sense of humor through-out our journey because PA can be vexing if we let it.

(1) The opinions expressed are the author's and do not infer endorsement by the U.S. Department of Health and Human Services

"Come in, TASA Control Center" (2)
"We read you, Captain, Are your readers' imagination switches on and wrist watches set for 4000 years?"
"I imagine so."

"Then get ready for countdown:
20th century
10th century
2000 BC
TIME OFF!"

Past

Put your camera on fast shutter as we race through Past.

The Biblical Period

Having never heard or read what the bible has to say about PA, I decided to find out for myself through a week of evenings spend tabulating words listed in the large subject index of the King James Version. The exercise was fun. While PA per se is not listed, here are some interesting findings:

o The bible refers to trait-related words like "meekness" twice as often (1,356 times) as to performance-related words like "deed" (632 times).

o The most frequently used qualifiers form this adjective title rating scale:

Lowest, Terrible	Lower, Fail	Low, Faulty, Unskillful	Acceptable, Fair	Great, High, Skillful	Greater, Higher, Excellent	Perfect Faultless, Greatest, Best, Highest

o Overall, the bible is more positive than negative in its judgments, with positive traits like "righteousness" and success-related references like "fruitfulness" outnumbering negative traits like "foolishness" and failure-related references like "unfruitful" 5 to 4 and 4 to 1 respectively.

o The judgments were consequential ones since they triggered these actions: punishment (51% of all actions), rewards (43%) and other (6%). Considering this finding in light of the one cited just above it, negative behavior was apparently punished more than positive behavior was rewarded.

(2) TASA - Time and Space Agency

Scripture adds much flavor to tabulations. Here is an example of the use of PA for promotion from Genesis 41 v 37,41: "The King said (to his officials), 'We will never find a better man than Joseph—' (and then) said to Joseph—'I now appoint you governor over all Egypt.'"

Much as I would like to, we can't linger for more. We must hurry on.

The Early Greeks

o Pythagoras, who taught that number was the essence of all things, may have been the first to introduce a numerical rating scale.

The Wei Dynasty

In the 3rd century AD, emperors of the Wei dynasty appointed an "Imperial Rater" to rate the performance of official family members. Sin Yu, a philosopher, was most happy about the process, saying that the highest ratings were given to favorites rather than to the meritorious. Sound familiar?

The Roman Empire

Thumbing through a large history book told me nothing about whether PA was responsible in any way for either the rise or fall of the Roman Empire. A colleague of mine, Jacques Jolie, who is a much better student of history than I am, gave me one tidbit from that period. It seems that Caesar had both a military and a civilian governor in Britain. Each would check on the other and report back to Rome. Relishing puns just as I do, Jacques noted that the reports may have been the first instance of peer "ratting."

Editor's Note: In a similar style, the author covered the historical period from 1500 to the present (skipping the dark ages).

Now

Our tour here includes side trips to the Land of Myth and Folly, Battlefield and the Land of MBR, and an overnighter at Ishu Inn. Some of what we will see had its start in Past.

Land of Myth and Folly

The IPMAAC tour guide will skim across 51 myths and follies located here and there in the literature and in practice. Since a myth or folly to me is someone else's belief or policy, you may not always agree with me. Second, some of the myths and follies do not involve PA per se, but do involve some related aspect of the broader process of performance management into which PA fits.

1. Person Appraisal. Traditional performance appraisal has been a misnomer. It's not performance appraisal at all, but instead

appraisal of the person's traits like "initiative," "perserverence" etc. Fortunately, I believe we are witnessing the demise of person appraisal through more enlighted employers, with or without the help of the courts.

2. Misdemeaning. "Performance is behavior." "Performance is results." I have read or heard many times these definitions of performance =, but neither is right in my opinion. Psychologists tend to use the first and business people, including management consultants, tend to use the second. My definition marries the two and is on the right side of the following, non-mathematical equation of human performance....:

Personal Factors + Situational Factors = Behaviors + Results
(Determinants of Performance) (Performance)

This simple equation has a lot of practical implications for managing performance, which we will see as we continue our tour.

3. Measurement myopia and sophomoric science. I dare say this because it takes one to know one. For more years of my career than I care to admit, I was myopic and trivial about PA. I saw PA as a measurement tool only and worked to foster more precise measurement of behaviors on the job.

   a. Forced-choice rating technique. Descriptive statements (usually behavioral translations of traits) are put into blocks of four statements each. Two statements are positive sounding, but only one, according to research, truly identifies successful job performance. Similarly, the other two are negative sounding, but only one is a true marker. Within each block, the hapless supervisor is forced to choose one statement that is most descriptive and one that is least descriptive of the employee. and any attempt to give meaningful feedback to the ratee is hopeless. No wonder the Army's use of this technique was fleeting. Yet, as recently as 1984, some U.S. companies were using it. (1)

   Cousins to this technique are forced rating distributions, "man-to-man" comparison ratings, straight rankings and, more recently, mixed standards scales. Common to them all is the objective of deflating ratings which we will meet at Ishu Inn. Suffice it to say here, I am in complete sympathy with the objective, but certainly not the methods.

   b. The format odyssey. History is cluttered with searches for the format with the best psychometric properties like resistance to leniency. An example is the Navy's adoption and rejection of 48 different kinds of efficiency ratings

from 1865-1956 (I facetiously call that episode the "ship of fools in search of the holy grail").
Two people did all of us a favor in 1981 when they reviewed 200 some studies and concluded that all of the different formats are about equally good (or equally bad depending on how myopic you are about measurement) in such properties.(2)

Their recommendations, though, are a mixed blessing: moratorium on format research (I certainly agree), more research on the statistical control of ratings (I definitely disagree since such control is what I call "numbers jumbo" and akin to techniques like forced rating distributions) and more research on the cognitive psychology of PA (another cul de sac as you can see at our next stop immediately below).

c. Brain picking. Cognitive research on PA is the study of the mental process of raters in the act of rating. So far, the payoff from such research is nil. (3) Just one example should show why.

The research goes like this. Subjects, as likely as not to be college sophomores, are given profiles of ratings of "paper people" on different behavioral dimensions. The subjects study on the profiles and then assign overall ratings to each of the overall ratings in conjunction with the dimensional ratings to see how well the former can be predicted by the latter and to figure out what weight or influence each dimension had on the overall rating.

You can see for yourself what is wrong with this research. Sophomores. Paper people. And a rating process that I certainly would not recommend because it focuses, usually exclusively, on behavioral dimensions, does not consider the role of weighting as a judgmental process in setting priorities during planning and presumes that overall ratings are derived by some complex mental process rather than more properly through a straight-foreword scoring procedure (adding up the products of the component weights and ratings) or through operational definitions (e.g., an overall rating of "outstanding" is defined and determined by a certain configuration of ratings on the components).

4. Fragmentation. PA needs to be seen and practiced as an integral part of a broader management function, yet too often is not.

Editor's Note: Several forms were mentioned.

a. Anniversary waltz. Another bad malady comes from organizations which schedule annual appraisals when employees' hiring anniversaries occur. What is foolish about it is that an organization does not manage the rest of its business on that schedule and thus cannot make (and fails

to appreciate the value of making) PA an integral part of business operations.

b. Segregated accountability. One example, "topless" PA, is executive immunity, and the excuse is that executives are held accountable in other ways and their jobs are too complex and dignified for PA. The problem with this excuse is, first, that executives should set an example, and second, that there is more to accountability than to the Board of Directors and share-holders. "Bottomless" PA is another example and refers to blue collar immunity. "Presumptive" PA is the third example and refers to the practice of presuming everyone is performing satisfactorily unless an Outstanding rating is requested or the employee is disciplined for poor performance. Presumptive PA is too presumptuous in my opinion.

c. MBO/PA divorce. This refers to the notion that the two are incompatible, that MBO is a very good planning process but much too idiosyncratic with its individually specific objectives to allow for equitable determination of merit pay allocations among individuals based on how well the objectives were achieved.

d. Split personality. This refers to the widely held belief that PA suffers from conflicting roles. The belief was perhaps best and first articulated by the late Douglas McGregor who felt that conventional PA forces supervisors to play the uncomfortable role of God or judge while facing the more modern and incompatible expectation of helping subordinates (5). His view was reinforced later by a General Electric study which seemed to show that appraisal meetings between supervisor and subordinate are dysfunctional if salary matters and performance improvement are both on the agenda (6). The literature has since been flooded with approving references to McGregor's view and/or the GE study and with recommendations to split out meetings or even to do separate appraisals for different purposes. I will explain briefly why I think the belief is a myth and the recommendations folly.

First, progress has overtaken McGregor's view. Conventional, or trait-oriented PA, is slowly but surely being replaced by approaches (e.g., MBO) which neither require supervisors to judge the person nor inhibit coaching.

Second, the GE study does not conclusively demonstrate the superiority of separating salary action from performance improvement discussions because the researchers confounded the separation with another experimental variable of high versus low employee participations in setting improvement goals, thus obscuring whatever effects the separation might have had. Further, there seems to have been an exaggerated, blanket emphasis on performance improvement. If the job is

getting done, searching for deficiencies and setting improvement goals can appear pointless and irritating to employees who neither need nor want improvement. I am not at all surprised that the researchers reported some managers tended to store improvement items so that there would be enough to talk about in the traditional, dual purpose meeting.

6. Birddogging. This refers to over-the-shoulder monitoring of employee performance. Daily diary keeping of employee behaviors and computer monitoring of outputs are examples. Now I have always believed that targeted follow-up prevents foul-up. And there is some evidence that effective performance managers do a better job of monitoring than ineffective managers (7). But birddogging is the antithesis of the more common sensical management by exception and self management and has been known to cause enough employee strife to attract the attention of the mass media (8).

7. Locked-in actions. This is what I call the locking of performance ratings to actions. An example of this is the mandating of awards of fixed amounts or above some minimum for given rating levels. Performance ratings need to be consequential because performance matters. At the same time, given the judgmental nature of ratings and the fact that performance is usually not the only legitimate consideration in any decision, a flexible link, not a lock, is needed between ratings and actions.

8. Perpetual marginals. This is the folly of allowing marginal performers to hover around marginality indefinitely.

9. Bending way over backwards. Cousin to the last folly, I mean here the unreasonable accommodation of substandard performers due to their personal circumstances. A real example is the case of the employer who lowered the acceptable production standard for a mentally handicapped worker to what would be a substandard level for non-handicapped workers in identical jobs. This is not good performance management, and would also be illegal if it occurred in the Federal government.

10. Two-letter managing. MR (like management-by-objectives) in which results or behaviors respectively are either undermanaged or not managed at all (9). Segregated accountability, which I put under the folly of "fragmentation," is also a form of two-letter managing.

I suspect you need a change of scenery now, so let's do a little sightseeing at the next scheduled stop. We will find some surprises and a quiz there.

Battlefield

This is strewn with court cases involving PA. It is advisable to follow a map. Two of the best maps were produced by Junior Feild and his colleagues

(10). The first was drawn from their sophisticated study of PA characteristics which differentiated between district court verdicts for and against the employer during the period 1965-1980. The second was a complete confirmation of the first using different court cases from 1982-1984. Characteristics which you might expect to influence the judges:

- o PA validity,
- o Rating errors and unreliability, and
- o Raters' qualifications and rater training

were found both times not to have unfluenced the judges. The characteristics which caused judges more often than not to rule against employers were these (in the order of their influence):

- o Trait-oriented instead of behaviorally-oriented PA,
- o Failure to give raters specific instructions on how to complete the appraisals,
- o Absence of a job analysis in developing the PA system, and
- o Failure to provide appraisal feedback to employees.

What encourages me about their studies is two-fold. First, finding yourself in court as an employer does not mean you automatically will lose. Second, to win means you only have to have been common sensical in your PA approach. You don't have to jump through rigorous hoops as some would lead you to believe.

Editor's Note: An extensive analysis and comments were made on court rulings.

Land of MBR

This side trip will be extremely short because we have made it before, and more detailed guides are available. MBR is a successful marriage of MBO and behavioral PA.

Following the MBR cycle, shown in Figure 1, is a good way to achieve "positive success." Note that PA is just one spoke of the wheel, and the least important spoke, too. As far as finding and steering performance goes, and there is not much more you could want, setting expectations is unbeatable.

MBR helps us see the double meanings of success and failure maybe in a new light. Please look at Figure 2 and see for yourself. If MBR is used properly, positive failure is never penalized like negative failure is. Actually, positive failure gets some credit. The most credit, of course, goes to positive success. And negative success? Well, in a competent and conscientious organization, "doing whatever is necessary" to succeed is no credo, either explicity or implicity.

FIGURE 1

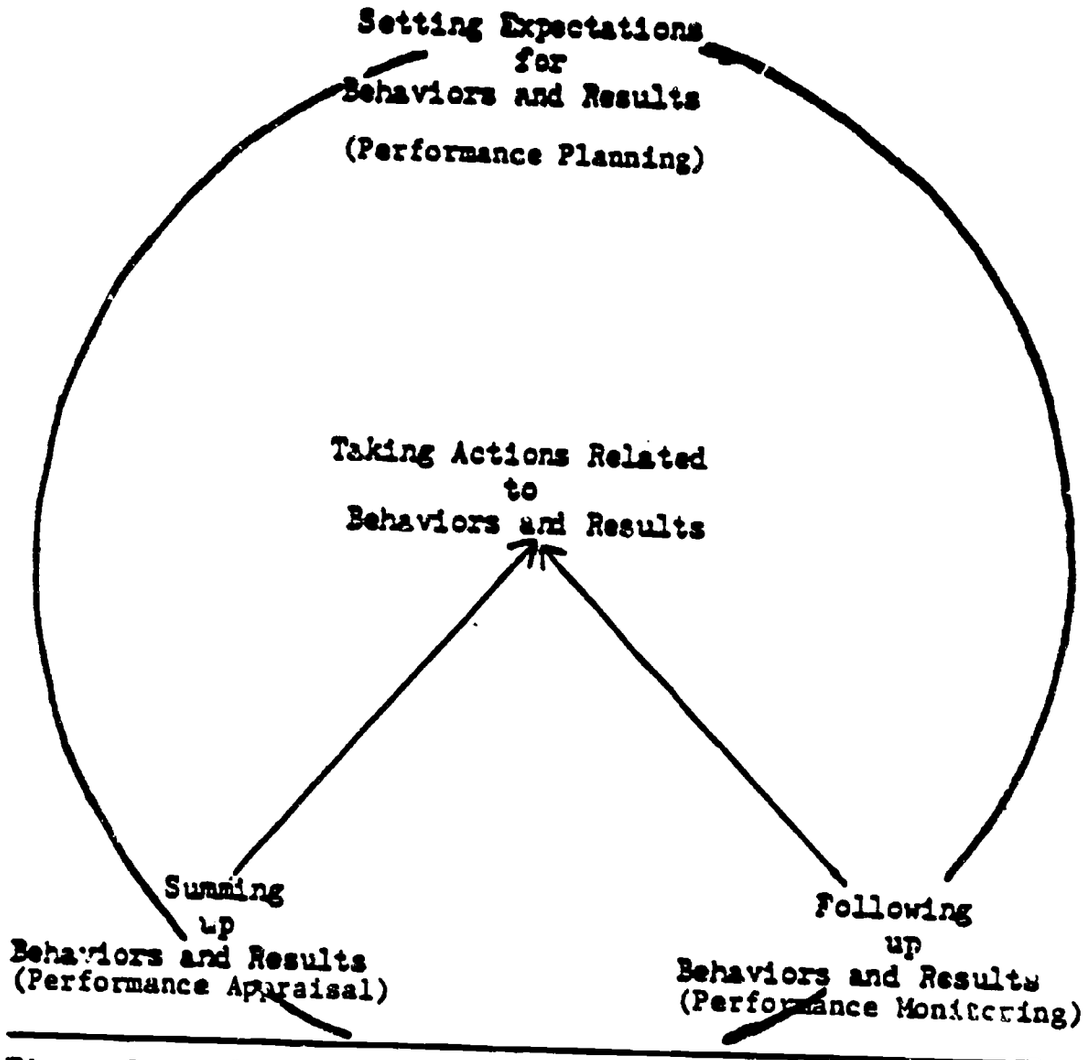


Figure 1. The MBR cycle to positive success.

Figure 2

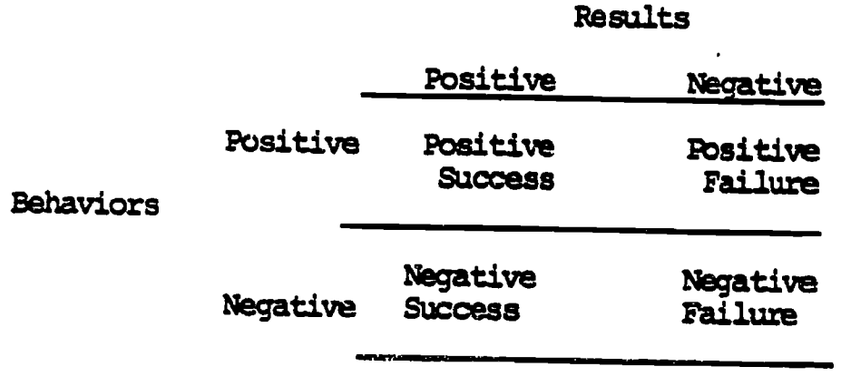


Figure 2. The double meanings of success an failure.

MBR can be fashioned in limitless ways. One of my favorites is a model we developed for senior executives and subordinate management ranks. It allows you to choose and explicity with the relative emphasis to be placed on the two parts of performance.

Ishu Inn

Let's stay overnight here before we head for Future. Don't expect much rest, though. From the 15 practical issues outlined in my IPMAAC tour guide, I have selected for a fireside discussion here the three which make me the most restless.

1. Sixteen elephants. This many pachyderms could not make a supervisor rate employees honestly according to a company president who was quoted in a recent newspaper article on civil service reform (11). His pillory appeared to be aimed at supervisors who do not manage the budgets from which merit payouts are drawn and thus see nothing to lose in giving inflated ratings.

   The issue here is not how to tell if supervisors are inflating ratings. The best way to tell is to look at the ratings, their documentation, and their association performance standards. A less direct way is to look at rating distributions. Suppose, for example, you saw this:

   70% of top management are rated "Outstanding"
   25% of rest of management are rated "Outstanding"
   10% of the general workforce are rated "Outstanding"

I presume you would think, as I do, that management, especially at the top, is mocking the meaning of "Outstanding." If I not unreasonably define "Outstanding" in part as representing "rare" performance, and then were to ask people how often they think something defined as "rare" occurs, most would probably say between five and ten percent of the time. Therefore, I would personally be suspicious of any percentage above 15 percent.

One option in solving the dilemma: There would be four levels available for rating managerial performance on the individual elements (expectations) in the performance plan. The levels would range from "Failed to meet the target" to "Exceeded the target." A fifth level, labelled either "Substantially exceeded" or in some places, "Outstanding," would be dropped, as would my belief that five levels are more natural for individual elements. By dropping the fifth level, supervisors would be relieved of the felt pressure to choose that level.

The ratings of a manager's performance on the individual elements would then be summarized, either by a scoring process (e.g., by summing the products of the elements' weights and ratings) or by operational definitions (e.g., "Exceeded the target" on most elements equals at least an "Excellent" summary rating). The summary rating would be put into one of four summary categories. All managers with ratings in the fourth category would be eligible for the reserved, rifth category of "Outstanding."

Criteria for distinguishing outstanding summary performance would be established through participative development by the managerial community. The criteria might define outstanding summary performance in terms of its dramatic and qualitative impact on organizational goals, its innovativeness, its complexity, its exemplary manner (behaviors), etc. One given criterion would be rarity of performance along with a policy guide saying rare performance would normally be expected to occur five to ten percent of the time. Another guide might say that at least the majority of the ratings on individual elements should be at the fourth level.

Managers who believed their performance met the criteria would nominate themselves for the fifth category. The nominations in effect would be self appraisals, the formal use of which I had earlier believed I could never advocate, thinking it would be a license for runaway ratings. But I have reread the literature and concluded that when self appraisals are not made anonymously, but instead forwarded to supervisors, you generally do not get exaggerated ratings (12). Modesty or the risk of embarrassment may help explain why this is so.

The immediate supervisor would be a conduit, or innocent bystander, through which the nominations would pass to a review committee. It would have the authority to pass judgment on the nominations in terms of the criteria and to pass out merit pay.

I cannot tell if you are rolling your eyes over this unorthodox option. If you are, one or more of the 17 other options might be more to your liking. Or if you have tried or thought of something unknown to me, please tell me.

2. Pay for performance. I sometimes wish this issue could be swept under the rug and forgotten.

I see the light in pay for performance with arguments for it like these: One, employees see the organization saying with its pocketbook that better performance matters. Two, pay determinants other than performance are not free from controversy either.

I feel the heat from arguments like these: One, money is not motivating, but getting less than Joe or Jane gets is mighty demotivating. Two, the prospect of a bonus turns one's attention from the task at hand to game playing (the assumption in this argument is that money does motivate, too well in fact).

If I had a choice, and were I free to try and come up with my own model approach, I think I would opt for pay for performance. In the meantime, I can only continue fretting about it.

3. Job-specific versus generic standards. A job-specific performance standard describes particular criteria for the performance of a particular individual in a particular position and is set when the performance plan is written to cover a particular performance period. Here is an example:

> "The multipurpose job analysis methodology must streamline and integrate the single-purpose procedures, be usable with all of the jobs in the organization, provide the information needed in the functions of (their names), be readily learnable, be acceptable to users, be pilot tested by (date) and ready for full use by (date)."

A generic performance standard describes general criteria for the performance of all people in similar positions, is pre-set and can be used for as many performance periods as the criteria remain relevant. Here is an example:

> "Creates an implementable solution to a routine problem."

Now, I am going to ask you a question about standards for the result part of performance. Would you say generic standards for results expected are more suitable for (a) executive, managerial and professional jobs or for (b) routine jobs? If you said (b), you agree with me and most if not all of the people on the IPMAAC time capsule journey.

Ready to leave the Inn? I am. There is not any more rest there for me. Besides, our capsule is waiting. Away we go!

Future

Oh, oh. Our windows have misted over. Can you forsee out. Here, let me try. I think I can barely foresee:

- o More people getting better at managing their own and other's performance.
- o More hybrids like MBR and less trait-oriented PA.
- o More widespread use of pay for performance, and the heck with the issue.
- o Continued litigation here and there, but fewer employer PA losses.

Touchdown

Well, we have landed safely. Anyone you walk away from is a safe one. Before you walk your fingers to the next article, please read my summary points:

- o PA has a long history
- o The history is chockfull of myths, follies, and battles worth some chuckles, shrugs, and chagrins.
- o Performance is both behaviors and results.
- o Managing behaviors and results gets you positive success.
- o PA is just one spoke in the performance management cycle.
- o The issues in PA and the rest of performance management are nettling, but manageable.

- o Remember to feed the elephants.

References

(1) Eichel, E. and H.E. Bender (1984). Performance Appraisal: A Study of Current Techniques, New York, American Management Association.

(2) Landy, F.Y. and J.L. Farr (1980). "Performance Ratings." Psychological Bulletin 87: 72-107/

(3) Banks, C.G. and K.R. Murphy (1985). "Toward Narrowing the Research-Practice Gap in Performance Appraisal." Personnel Psychology 38: 335-345. See also: Ilgen, D.R. and J.L. Favero. (1985). "Limits in Generalization from Psychological Research to Performance Appraisal Processes." Academy of Management Review 10: 311-321.

(4) Brumback, G.B. (1978). "Toward a New Theory and System of Performance Evaluation: A Standardized MBO Approach." Public Personnel Management 7: 205-211; Brumback, G.B. (1981). "Revisiting an Approach to Managing Behaviors and Results." Public Personnel Management 10: 270-277; and Brumback, G.B. and T.S. McFee (1982). "From MBO to MBR." Public Administration Review 42: 363-371.

(5) McGregor, D. (1957). "An Uneasy Look at Performance Appraisal." Harvard Business Review 35: 89-94.

(6) H.H. Meyer, E. Kay and J.R.P. French, Jr. (1965). "Split Roles in Performance Appraisal." Harvard Business Review 43: 123-129.
(7) Komaki, J. (1986). "Effectively supervising others: Documented day-to-day interactions." Invited address to the Personnel Testing Council of Metropolitan Washington, April.
(8) Perl, P. (1984). "Monitoring by Computers Sparks Employee Concerns." The Washington Post September 2.
(9) Brumback and McFee, op cit.
(10) Feild, H.S. and W.H. Holley (1982). "The Relationship of Performance Appraisal System Characteristics to Verdicts in Selected Employment Discrimination Cases." Academy of Management Journal 25: 392-406; Feild, H.S. and D.T. Thompson (1984). "Study of Court Decisions in Cases Involving Employee Performance Appraisal Systems." The Daily Labor Report December 26.
(11) Havemann, J. (1986). "Civil Service Reform Remains in Vogue." June 15: The Washington Post, p. A6.
(12) See, e.g., H.H. Meyer (1980). "Self-Appraisal of Job Performance." Personnel Psychology 33: 291-295.

Note: The author indicated he would gladly make additional materials available to those requesting them.

* * *

Bootstrapping Drafters on the Bay
Summary

Thomas A. Tyler, Ph.D., Merit Employment Assessment Services, Inc.
Flossmoor, Illinois

I. Introduction.

A job-analysis was performed on three classes of drafting positions (Drafters) for the City and County of San Francisco (on the Bay). These positions were Civil Engineering Assistant I (Position 5360), Civil Engineering Assistant II (Position 5362), and Civil Engineering Associate I (Position 5364).

This analysis revealed that the positions varied from a beginning level "board" position to an advanced, nearly-professional level civil engineering position. The variety of work varied through several public works departments from the water supply in the Sierras, to the airport, to the Muni railroad, to the Assessor's Office, and beyond.

Another complication was that some candidates would be eligible to take the examination for two of three levels; and some candidates would be eligible to take all three examinations. The final consideration was that a large proportion of the candidates were Asian-American resident-alien Asians.

Although one could argue that there were sufficient differences in these various positions to justify a number of different examinations, the fact existed that each of these employees is administratively transferable between any of the departments. Furthermore, a common core of basic skills existed at each of the three levels. To cover the diversity of the positions it was decided to measure the knowledge, skills, and abilities with a wide variety of procedures. Thus, it was necessary to develop an objective and valid means of combining the scores from these diverse procedures for a final eligible list. For this purpose it was decided to use a multiple-regression procedure (Bootstrapping) to derive weights for each of the components of the examinations.

The key element in the bootstrapping procedure is the use of content-experts to form a "selection" panel. This panel reviews all of the available information on the candidates and assigns a subjective rating to each candidate. This subjective rating is then used as a "criterion" rating to determine the regression weights to be applied to the "predictors" (various objective scores from the tests). Although the panel may review application forms, experience data, etc., the final regression equation involves only the objective test scores and it is therefore objective in total and consistent with civil service procedures.

Research performed by Dawes (1971) has indicated that bootstrapped scores can be much more valid than the judgments they were derived from; and Tyler (1980) has argued that bootstrapped validity is theoretical superior in many ways to the traditional empirical validity models.

II. Examination Materials.

Twenty-three KSA's were identified for measurement in the job-analysis. Avoiding excessive detail, the following instruments were developed:

A. A different, but overlapping, multiple choice exam for each level.

B. A single checking test and single filing test (both speeded) common for all levels.

C. A drawing performance test for the lowest level, and a second drawing performance test for the top two levels. These perform-

ance tests were pre-printed on drafting paper with a series of exercises to be performed (e.g., lettering, layout, scaling, etc.)

D. A writing sample for the upper two positions requiring a written report based on simulated information.

E. A structured oral for the entry level based on a critique of a badly-drawn blueprint. For the upper two levels the candidates were to critique this drawing from a supervisory perspective.

F. A variety of instructions to candidates and raters including scoring templates for the performance test and an elaborate Study Guide for the candidates.

This variety of tests, candidate instructions, rater guides, scoring templates, etc., was so large that a "catalogue" was prepared to keep the procedure manageable.

III. Rating Panel.

After all of the objective and performance tests had been administered and scored a panel of two supervisors was formed for each of the three positions. Training of the raters included actual administration of the written exams, review of all testing materials, explanation of standard scores, and the usual training in the use of rating forms. Each panel was presented with a standardized profile of test scores for each candidate. In addition, the panel was provided with each candidate's application form which included educational background and experience, and the candidate drawings. Each member of the panel assigned a rating to each candidate. Candidate data was anonymous.

IV. Analysis and Result.

Multiple regression was performed between the several tests and the average of the two ratings made at each level or class. The program used was REGRESS from Human Systems Dynamics of Northbridge, California. There was some concern that the written material might discriminate against the Asian-surnamed candidates. For this reason, an English grammar test was included in the written test. If necessary, separate regression analysis would have been made for each ethnic group. However, the Asian-surnamed candidates performed slightly better on the English grammar subtest than the remainder of the candidates so it was decided that language was not a handicap and a single analysis was indicated. The regression weights, multiple R and significance are given for the three classes in Table I. All scores were converted to T-Scores (mean = 50, standard deviation = 10) before the regression analysis so that the values of the regression weights can be rather directly compared.

Inspection of Table I indicates a large and statistically significant multiple - R ranging from .76 to .82. This would be expected from the design but does indicate that the regression procedures rather faithfully model the human judgments. The major contributor in each class is the drawing performance test, which seems reasonable for positions which are defined as drawing or drafting positions. Another reason for this large contribution might have been the high quality of the performance test, including a careful standardization of the scoring procedure. One could speculate on the contributions of the other tests but with the small sample size compared to the number of variables such speculation is of little value. After this analysis, the predicted scores (weighted composites) were converted from the rating point scale (0-5) to a 700 to 1000 point scale used by the Civil Service Commission for eligible lists. Inspection of the lists indicated no adverse impact on the Asian-surnamed candidates at any of the tested cutting points at any of the three levels.

Table I
Regression Analysis

Test	Ass't. I	Ass't. II	Assoc. I
Checking (a)	.0077	.0416*	.0111
Filing (a)	.0119	-.0013	.0302*
Multiple Choice (b)	.0368*	.0172	.0394**
Drawing (c)	.0996***	.0734***	.0699***
Writing (d)	N/A	.0272	.0112
Oral (c)	-.0008	.0337*	.0159
Constant	-4.2400	-5.5783	-4.6215
Sample Size	50	53	47
Multiple R	.7902***	.7562***	.8162***

* Significant at .05
** Significant at .01
*** Significant at .001

(a) Same test at all three levels
(b) Different test at all three levels
(c) One test at lowest level, different
(d) Not administered at lowest level

References

Dawes, R.M. A case study of graduate admissions. Application of three principles of human decision making. American Psychologist, 1971, 26, 180-188.

Tyler, Thomas A. - Bootstrapping - A Primer. Unpublished Monograph 1980.

References

Dawes, R.M. A case study of graduate admissions. Application of three principles of human decision making. American Psychologist, 1971, 26, 180-188.

Tyler, Thomas A. - Bootstrapping - A Primer. Unpublished Monograph 1980.

* * *

A MINI-WORKSHOP

Passing Point Methodology

Susan Christopher, State of Wisconsin Department of Employment Relations

Barbara Showers, State of Wisconsin Department of Regulation and Licensing

This workshop considered how to determine passing scores for either civil service tests - which are primarily used as ranking procedures, or licensing - which are used to establish whether an individual meets minimal qualifications for entry into an occupation or profession.

Coverage of Workshop

- Introduction and Overview
- Factors Affecting Passing-Point Determination
- Traditional Methods of Setting Passing Points
- Competency-Based Methods

  A. Angoff
  B. Nedelsky
  C. Application
  D. Discussion

- Summary

I. Overview

It is important to point out a few things about passing points. Once you have administered a test, it is necessary to decide who passes and who does not.

1. There is no one right way or single method for setting a passing point. The factors in each situation may affect where the point is set.

2. It is always a judgmental process - whether you rely on your opinion as the test expert, the opinion of subject matter experts, or the statistical characteristics based on one or more administrations of the test.

3. What is necessary is to find/determine a defensible passing point - not only a legally defensible one, but a defensible one because it is competency-based. The objective is to use information relevant to the situation to produce a defensible, fair decision. Passing points can well be looked at in terms of risk:

    1. Legal risk - can I defend where I set the passing point
    2. Risk to management in hiring an incompetent or not having a competent person available because I failed them.

II. Factors Affecting Passing Point Determination

However you set your passing points, it is necessary to look at defensibility. There are some factors you can consider which will help you in your decision. Several factors which would impact on where the passing point would be set can be considered prior to administration - and several other factors must be considered after the test is administered. For each factor there should be considerations of: a) what is the concept? b) how does the factor affect the passing point? c) are there differences in the effect of the factor for different uses of the test, e.g., civil service hiring - licensing?

III. Validity of Recommended Passing Point

If the test is not job-related or is only marginal so, best to be very cautious in setting pass points. If the test is job-related, some of the same methods that are used to validate the test can be used to set the passing point. Here, pass point validity refers to the relationship of the pass point to minimally acceptable job performance. For example, if you have criterion-validation data, you may use it to identify the test score which predicts acceptable job performance. If you are using a content validation strategy, you may use job experts to judge the best passing score. If you are evaluating whether the passing point is appropriately set, one of many things to look for is evidence of its relationship to job performance. This is a fundamental requirement for defensibility. Other factors can be used to adjust the pass point, but its basic meaning is to separate competent from incompetent. You can't let the other factors take you too far from this concept.

Raters, that is job experts or subject matter experts, are polled for their opinion of the passing point. The validity of this recommendation depends on whether proper procedures were used and whether an adequate number of raters were used and whether the raters or subject matter experts were representative of all the jobs for which this test was or will be administered.

Reliability/Standard Error of Measurement (SEM)

Once the test is administered, the reliability and the standard error of measurement are determined. These two statistics indicate score accuracy. The more accuracy, i.e., the smaller the SEM, the more likely a person's observed score represents his/her true score. The standard error, expressed in test score units, reflects the range of scores in which the candidate's "true" score lies, e.g., if observed score is 70 and SEM is 5, then the "true" score is between 65 and 75 about 68% of the time, and between 60 and 80 about 98% of the time. Notice that 60 to 80 is a large range of uncertainty.

When you are attempting to set a precise passing score, this range of uncertainty can be a problem. It affects the interpretation of the pass--fail point as a clear indicator of competent or incompetent.

When setting the passing point, there are a number of philosophies which attempt to deal with this uncertainty.

Suppose we are given a job expert recommendation to set the passing point at 70 points:

> Philosophy #1: If this is a job with substantial risk to the public, we may want to assure that no incompetents are hired, so we raise the passing score 2 SEMs to avoid the possibility of hiring someone who's "true" score is below 70, but whose observed score through error is above 70 [shown on flipchart]. This may fail a number of competent candidates, but we feel the risk to public health outweighs the interests of these candidates [National Nursing licensure exam does this].
>
> Philosophy #2: This is a job where all candidates are ranked, the low scoring candidates are unlikely to be considered for hire, and/or our public sector employment philosophy requires that the benefit of the doubt be given to the candidate. Then, we might lower the pass point up to 2 SEM (or even 3) to be sure to include all candidates whose "true" score maybe at least 70, but whose observed score, through error, is less. This may pass a number of incompetent candidates, but we feel the benefits to the candidate outweigh the risk to the public.
>
> In both cases, the link to the job related pass score recommendation is maintained by adjusting the pass point within limits of possible error.
>
> Philosophy #3: Give no benefit of doubts either way and accept the job experts' recommendation. The philosophy here is that error can occur in either direction and the candidate's observed score is our best estimate of the true score. While the candidate can argue a score below passing is due to error, management can equally argue that it is already higher then it ought to be due to error, and may actually be lower than reported.

IV. Adverse Impact

Uniform Guidelines definition: pass rate of one group is less than 80% of pass rate of another. Usually minority is lower. Can also be statistically significant differences in pass rate.

Adverse impact, if it exists, frequently hampers the ability of "managers" on reaching (meeting) their affirmative action goals; additionally, evidence of adverse impact places a burden on the test user under EEOC guidelines to assure the validity of the test. Using a methodology like lowering the recommended passing point by SEM units may allow affirmative hires and may reduce the adverse impact.

V. Past Passing Points and the Number of Times a Test May be Used Again

Whatever the methodology used in setting points, it is important to be consistent over time and across administrations. Obviously, changes in the passing point are difficult to defend.

VI. Vacancies

How many vacancies are to be filled from this pool of people is an important factor to consider is setting the passing point for civil service tests. (It may not be so important for licensing examinations, however.)

Obviously, the more vacancies in relation to the number of qualified applicants, the more likely you will have to consider lowering the recommended passing point by some SEM units. Similarly, if you have only a few anticipated vacancies in relation to the number of qualified applicants, the more likely you will retain the recommended passing point or possibly raise the passing point by some SEM units.

- In the case of civil service testing - passing more individuals than you need is not a sin - people do not like to be called ineligible or failures and there may be no purpose served in raising the passing point.

VII. Gaps in the Distribution

Of all the factors, this is probably least important of all - It may be useful in a one-time administration, since you can increase the passing point reliability if you set the passing point in a gap since there are no scores immediately next to the passing score.

VIII Some of the More Traditional Methods of Setting Passing Points

There are four traditional methods which come to mind: 1) percentages, 2) norm-referenced, 3) gaps, 4) numbers of people.

1. Percentages simply means setting the passing score at some

arbitrary percentage - generally that percentage is 70%. (e.g., if you had 90 items on a test - the passing score would be 63 items.)

2. Second method is norm-referenced and usually looks something like: minus 1 (or more) standard deviations. If your distribution is normal, then you would be passing 84% of the candidates if you set your passing point at 1 standard deviation below the mean.

3. Another approach - is look for gaps in the score distribution - helps reliability of that passing point if no individual is actually on that point.
4. Finally - look at numbers of people, what percent of the group - do you want to pass or do you want to fail?

Example of Data for Comparisons of Traditional Methods.

Number of candidates:	20
Total possible raw score:	85
Mean:	73.35
Standard Deviation:	5.34

Distribution
83
80
79
79
78
78
77
76
75
74
72
71
71
71
70
69
68
67
65
64

Comparison

% of (70%) =	59.5
Norm-Referenced Mean =	73.35
Mean - 1 s.d. =	68.01
Mean - 2 s.d. =	62.67

Gaps 73 , 66

Numbers of People

Pass 50% = 74 (or 73; (10 people)

Pass 20% = 79 (4 people)

In comparative studies, Nedelsky tends to give lower pass points then Angoff on same items.

IX. Methods of Item or Question Analysis

Angoff method is very simple - ask judges to identify for each item "What percent of minimally competent new employees would get this item correct?" Pick any percent, or give choices. Average results for each item, and for all items used in test to get recommended passing score.

PROs: Competency related, easy to understand, cheap
CONs: Judgement can be questioned, SMEs must be representative, reliability problems (Fight rater bias toward traditional #'s, e.g., 70%)

Nedelsky method is more complex - ask judges to eliminate the distractors that the minimally competent candidate would eliminate. Then compute from the choices remaining the probability of the candidate guessing the right answer.

PROs: Competency related, may be more precise than Angoff, simulates candidate test-taking behavior, reduces rater bias toward 70%.
CONs: Same as Angoff re: judgments, representatives and SMEs, also more difficult to explain.

OUTLINE OF METHODS COVERED FOR DETERMINING PASSING POINTS

METHODS	WHEN	HOW	PROS	CONS
I. Traditional				
A. Absolute				
% correct	Last choice; when validation not necessary or when the test difficulty can be adjusted to fit a validated standard.	Calculate: Final Score = Raw Score / Total Possible	Easy to calculate; totally objective; may have traditional usage/acceptability.	Not job or test related. Not fair.
B. Norm Referenced	Ranking is important; reasonable prior assurance of general competency of group; large group of candidates takes test.	Use descriptive statistics. Usually the mean minus one or more Standard Deviation.	Relatively easy to calculate. Assures that the best of the group passed and the worst failed.	Related to group performance, not job performance.
Group performance on test: Mean - SD, etc.				
II. Empirical				
A. Related to existing on-the-job performance measurement.	Criterion-validated test. Test scores statistically related to job performance. Large job classes with large number of hires and large number of incumbents.	Regression equation, expectancy table.	Clear job relatedness. Likelihood of success is known.	Not feasible for small job classes. Costly, time consuming for large classes.
B. Contrasting Groups.	Access to two clearly-defined groups where one is known to be qualified and one is known not to be qualified.	Compare test performance of two groups; one qualified, the other definitely doesn't know the subject matter.	Direct measure, should find passing point that separates groups.	Expensive. Relies on volunteers who may not be well motivated.

METHODS	WHEN	HOW	PROS	CONS
III. Judgmental Subject Matter Expert				
Evaluation of test performance based on test content.	Content validated tests, prior to test.	SME judgment.	Competency related; relatively cheap; credibility.	SME judgment can be questioned. Depends on represenative sample of SME's. Problems with reliability of ratings. Different methods produce different results
A. Angoff Method	As above.	Raters judge % of minimally competent who will be successful for each item.	As above	As above
B. Nedelsky	As above.	Raters identify the distractors that minimally competent candidates would elimate.	As above.	As above.

POSTER SESSION

The Effects of Sex-Role Stereotypes on Personnel Decisions

Edward H. Hernandez

University of California at Long Beach

Within the organizational context it is necessary to conceptualize sex discrimination as having two components: access discrimination and treatment discrimination (Terborg & Illgen, 1975). Access discrimination refers to non-job related limitations placed on a subgroup at the time a position is filled. Rejection of applicants for nonjob-related reasons, lower starting salaries, closure of higher skill level jobs, and failure to recruit applicants for certain positions from the subgroup population represent some forms of access discrimination (Levitin, Quinn & Staines, 1971). Treatment discrimination refers to differential treatment of subgroup members once they have gained access into the organization. Slower rates of promotion, lower and less frequent raises, less training opportunities, assignment to less attractive or less challenging tasks, etc., represent some forms of treatment discrimination.

With respect to traditionally masculine occupations, access sex discrimination has been demonstrated repeatedly in employee selection (Fidell, 1970; Jones, 1970; Shaw, 1972; Wiback, Dipboye, & Frompkin, 1975; Cash, Gillen, & Burns, 1977; Terbog & Illgen, 1975). Women often have encountered various forms of discrimination such as the withholding of rewards, facilities, or opportunities which are legitimately deserved (Terbog & Illgen, 1975). Another possible explanation for these findings is given by Broverman, Vogel, Clarkson & Rosenkrantz (1972) who found that competence is considered stereotypical of men, but is not generally expected of women. Thus, to "protect" the organization, administrators allegedly resort to a pattern of exclusion in selection which bars women from the more challenging roles or places them at a disadvantage when they do achieve these roles (Rosen & Jerdee, 1974).

Evidence also exists which indicates that women are being discriminated against on treatment variables. Discrimination has been reported in promotions (Bryce, 1970; Day & Stogdill, 1972; Rosen & Jerdee, 1974), employee utilization (Knotz, 1970), employee development (Rosen & Jerdee, 1974), and pay allocation (Levitin, Quinn, & Staines, 1971).

With respect to traditionally feminine occupations, access sex discrimination has been demonstrated by Cash, Gillen & Burns (1977) where male applicants are discriminated against when applying for traditionally female jobs.

Men have also been demonstrated to be discriminated against on treatment variables. Rosen & Jerdee (1974) and Rosen, Jerdee & Prestwich (1975) found that any intrusion of family or other personal considerations may be viewed more unfavorably for men then for women.

On the basis of commonly alleged stereotypes for males and females it is hypothesized that subjects would tend to discriminate against females in important decisions involving promotion, hiring (into neutral, male dominated, and complex occupations), development, allocation of responsibility, and punishment. It was also hypothesized that subjects would tend to discriminate against males in decisions involving competing role demands stemming from family or other personal circumstances, and in hiring decisions when applying for traditionally feminine jobs.

METHOD

Subjects

A questionnaire was given to 42 male and 59 female undergraduate students attending introductory psychology classes at California State University at Long Beach. Their average age was 19.6. 75.2% of the students in the sample stated that they are presently employed and those employed stated that they work an average of 21.9 hours per week.

Procedure

In order to reduce the potential effects of a social desirability response set due to direct questions regarding sex discrimination and sex-role stereotypes, a survey-experiment was developed in the form of "in-basket" decision-making tasks. Students in the sample were asked to read several incidents in the form of letters and memorandums depicting various organizational problems.

The in-basket format was used to increase the realism for making managerial decisions. Also, real stationary from actual organizations was used for memorandums. Finally, a between group design was chosen for this experiment where most administrative decisions deal with only one employee. It is assumed that when subjects encounter a choice between a male and a female for personnel decisions, the issue of discrimination becomes more obvious.

Hiring into position of Personnel Officer: Sex of candidate and complexity of Job: This item was in the form of memorandum requesting a decision on the hiring of a candidate to the position of Personnel Officer. The memorandum was written in four versions so as to manipulate the variables of sex of candidate and complexity of the job. The job was described as either a very-complex upper-management position given much responsibility or a moderately easy supervisory position given few substantial responsibilities. Subjects were told that the position had become vacant since the last person to hold it had retired.

Attached to each memorandum was a resume of qualifications of the candidate. Half of the resumes had the name John Williams as the candidate and the other half had Jane Williams. The major dependent variables were (a) a rating of the applicant's qualifications on a 9-point scale from "very unqualified" to "very qualified", (b) a rating of the subject's expectations of the applicant's future performance on a 9-point scale from "very unsuc-

cessful to very successful", (c) a rating of the subject's recommendation of the applicant on a 9-point scale from "strongly recommend not hiring to strongly recommend hiring", and (d) a rating of the applicant's overall employment potential on a 9-point scale from "low potential to high potential." The 2x2 between group experimental design for this item included 2 factors (Sex of applicant x Complexity of Job).

Employee Development: Sex of applicant x Cost of Development: This item was in the form of a memorandum asking for subject's opinions about sending an employee to a class on strategic marketing management. On half of the memos, subjects were asked to rate sending a female employee (Amy Davis). On the other half of the memos, subjects were asked to rate sending a male employee (Tom Davis). For both the male and female versions, half of the subjects were asked to rate sending the employee to a $140 Extended Education class at U.C.L.A. on strategic marketing management, and the other half to a $6000 Executive Education Program at Harvard University on strategic marketing management. The memo states that the employee has been the assistant to the Marketing Director for the last 4 years and has a degree in Marketing. Thus, a 2x2, between-group design was used (Sex of Employee x Cost of Development) for this item.

On the memos were two 9-point rating scales asking the subjects to (a) give their recommendations regarding sending the employee from "strongly recommend not sending to strongly recommend sending", and (b) stating how much they feel the employee would benefit from the program from "will not benefit much from this program to will greatly benefit from this program". Also, subjects were asked if someone else should be found to send rather than the employee on the memo. It was expected that male employees would more likely be sent to the high cost development program and that female employees would more likely be sent the low cost development program.

Salary for Promotion: In this item subjects were asked to read a memorandum describing a situation where an employee, either male or female, is being promoted to the position of Manager of Production. Subjects are told that this employee was being paid $27,000 on his/her old salary. Finally, subjects are asked to give a dollar amount from $0 to $10000 indicating amount for a raise the employee should receive. It is expected that the male employees will receive a higher raise than the female employees.

RESULTS

Hiring into position of Personnel Officer: Sex of Candidate and complexity of job: Table 1 indicates the mean ratings for hiring of male and female employees into both simple and complex jobs. With regards to the perceived qualification of the employee on the low and high complexity jobs, the differences between the male and female ratings were not significant. However, for the high complexity job the male candidate received a higher rating and for the low complexity job the female candidate received the higher rating.

With regards to the expectations of the applicants' future success for both the low and high complexity jobs, the differences between the male and

female candidates were not significant. However, for the high complexity job the male candidate received a higher rating and for the low complexity job the female candidate received the higher rating.

With regards to the expectations of the hiring recommendation there was a significant difference between the male and female scores for the high complexity job (p less than .05). Males were more likely than females to receive a more favorable recommendation to be hired into the high complexity job. For the low complexity job there was no significant difference between the male and female ratings. However, the females received the higher score for the low complexity job.

With regard to the rating for overall employee potential with both high and low complexity jobs, there were no significant differences. However, the male candidate received higher ratings for the high complexity job and the female candidate received the higher ratings for the low complexity job.

When collapsing the previous four dependent variables together to get an overall employability rating there is a significant difference between the male and female rating for the high complexity job (p less than .01). Although the difference was not significant, the female candidate received the higher rating than the male candidate for the low complexity job.

Employee Development: Sex of Applicant x Cost of Development: Table 2 shows subjects ratings of male and female employees for low and high costing development.

Male applicants were more likely than female applicants to be sent to the low cost development program. Also, subjects considered male employees as benefitting more than female employees from the low cost development (p less than .06). When collapsing the two dependent variables for the low cost development together it is found that males are significantly more likely than females to be sent to the low cost development program (p less than .025).

For the high cost development, contrary to expectations, female employees are more likely than male employees to be sent. Also, female employees were perceived as being more likely to benefit from the high cost development program (p less than .20). When collapsing the two dependent variables for the high cost development together it was found that female employees were more likely than male employees to be sent to the high cost development program (p less than .10 2-tailed; p less than .05 1-tailed). Figure 3 demonstrates the interaction between sex of employee and cost of development.

Salary for Promotion: There was no significant difference with the salary increase given to the male or female employees.

DISCUSSION

Results from this experiment confirm the hypothesis that males would be looked upon more favorably for more complex occupations and that females would be looked upon more favorably for less complex occupations. These results are very similar to those found in Rosen & Jerdee (1974) demonstrating a different treatment by sex in promoting male employees into more complex upper-management occupations. This study also investigated what may become an increasingly serious role conflict for male and female employees; the conflict between career and family responsibilities. In both Rosen & Jerdee (1974) and Rosen, Jerdee & Prestwich (1975) it was found that it is considered significantly more appropriate for a female to ask for time off from work to take care of children. However, my findings show that it is considered significantly more appropriate for males to take the time off. These differences may be due to the fact that in both the other studies only male managers were used as subjects. Despite objectively equivalent qualifications, job applicants may encounter different employment opportunities that are dependent upon their sex and sex-role characteristics of the opportunities they seek. Bias continues to operate against out-of-role positions for both males and females. Among occupations of low to moderate prestige and skill considered in this experiment, sexist effects have a clear influence on the opportunities for employment.

Numerous studies used to formulate theories of sex-role stereotyping have used exclusively male subjects (e.g., Rosen & Jerdee, 1974; Rosen, Jerdee & Prestwich, 1975). Future research should replicate these studies using female studies. Future research should replicate these studies using both male and female managers. This is mainly due to the increased number of females in managerial ranks since the time these studies were conducted. Also, the different findings between these and this study on similar measures may indicate differences between the sex of subjects when measuring sex-role stereotypes.

Table 1
Mean Ratings for Hiring of Male and Female Employees into Simple and Complex Jobs.

	Job Complexity			
	Low		High	
Perceived Qualifications				
Males	6.79		5.90	
Females	7.08		5.36	
	t[a]	<1	t[b]	1.08
Expectations of Applicant's Future Success				
Males	7.39		6.70	
Females	7.46		6.33	
	t[a]	<1	t[b]	<1
Hiring Recommendation				
Males	7.07		6.34	
Females	7.58		5.48	
	t[a]	<1	t[b]	1.73*
Overall Employment Potential				
Males	7.00		6.82	
Females	7.58		6.20	
	t[a]	1.09	t[b]	1.28
Overall Employee Rating (previous four factors collapsed.)				
Males	7.06		6.44	
Females	7.34		5.83	
	t[a]	1.25	t[d]	2.47**

[a]df = 50. [b]df = 48. [c]df = 206. [d]df = 197. *p<.05 **p<.01

Table 2
Employee Development: Mean Ratings by Sex of Employee and Cost of Development.

	Sex of Employee		
	Male	Female	t
Low Cost Development			
Recommendations for Sending to Program	7.44	6.79	1.21
Perceived Benefit Attained by Going	7.44	6.43	1.99*
Both Variables Collapsed Together	7.44	6.60	2.27**
High Cost Development			
Recommendations for Sending to Program	7.00	7.50	1.02
Perceived Benefit Attained by Going	7.27	8.00	1.59
Both Variables Collapsed Together	7.14	7.77	1.89***

*p<.05 df = 52. **p<.025 df = 105. ***p<.05 1tail p<.10 2tail df = 93.

References and Additional Readings

Berman, E., Sacks, S., & Lief, H. (1975). The two-professional marriage: A new conflict syndrome. Journal of Sex and Marital Therapy, 1:5, 242-253.

Broverman, I.K., Vogel, S.R., Broverman, D.M., Clarkson, F.E., & Rosenkrantz, P.S. (1972). Sex-role stereotypes: A current appraisal. Journal of Social Issues, 28, 59-78.

Cash, T.F., Gillen, B., & Burns, D.S. (1977). Sexism and "beautyism" in personnel consultant decision making. Journal of Applied Psychology, 62, 301-310.

Clark, R.A., Nye, F.I., & Gecas, V. (1978). Husband's work involvement and marital role performance. Journal of Marriage and the Family, February, 9-12.

Day, D.R., & Stogdill, R.M. (1972). Leader behavior of male and female supervisors: A comparative study. Personnel Psychology, 25, 353-360.

Dipboye, R.L., Fremkin, H.L. & Wilback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. Applied Psychology, 60, 39-43.

Ferber, M. & Huber, J. (1979). Husbands, wives, and careers. Journal of Marriage and the Family, May, 315-325.

Fidell, L.S. (1970). Emperical verification of hiring practices in psychology. American Psychologist, 25, 1094-1098.

Gove, W.R. & Geerken, M.R. (1977). The effects of children and employment on the mental health of married men and women. Social Forces, 56:1, 66-76.

Gutek, B.A., Nakamura, C.Y. & Nieva, V.A. (1981). The interdependence of work and family riles. Journal of Occupational Behavior, 2, 1-16.

Hopkins, J. & White, P. (1978). The dual-career couple: constraints and supports. The Family Coordinator, July, 253-259.

Huser, W.R., & Grant, C.W. (1978). A study of husbands and wives from dual-career and traditional-career families. Psychology of Women Quarterly, 3, 78-89.

Johnson, C.L. & Johnson, F.A. (1977). Attitudes toward parenting in dual-career families. American Journal of Psychiatry, 134:4, 391-395.

Jones, R.H. (1970). Sex prejudice: Effects on the inferential process of judging hireability. Dissertation Abstracts, 31, 1013A.

Kaley, M. (1971). Attitudes toward the dual-role of the married professional women. American Psychologist, 3:26, 301-307.

Katz, M.H. & Piotrkowski, C.S. (1983). Correlates of family role strain among employed black women. Family Relations, 32, 331-339.

Keith, P.M. & Schafer, Robert B. (1980). Role strain and depression in two-job families. Family Relations, 29, 483-488.

Kootz, E.D. (1970). Women's bureau looks to the future. Monthly Labor Review, 93, 309.

Levitin, T., Quinn, R.P. & Staines, G.L. (1971). Sex discrimination against the American working women. American Behavioral Scientist. 15, 238-254.

Lewis, R.A. & Pleck, J.H. (1979). Men's roles in the family. The Family Coordinator, October, 429-432.

Murstein, B.I. & Williams, P.D. (1983). Sex roles and marriage adjustment. Small Group Behavior, 14:1, 77-94.

Rapoport, R. & Rapoport, R.N. (1971). Further considerations on the dual-career family. Human Relations, 24, 519-533.
Rosen, B. & Jerdee, T.H. (1974). Influence of sex-role stereotypes on personnel decisions. Journal of Applied Psychology, 59:1, 9-14.
Schein, V.E. (1973). The relationship between sex-role stereotypes and requisite management characteristics. Journal of Applied Psychology, 57, 95-100.
Shaw, E.A. (1972). Differential impact of negative stereotyping in employee selection. Personnel Psychology, 25, 333-338.
Terborg, H. & Illgen, D. (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. Organizational Behavior and Human Performance, 13, 352-376.
Waite, L.J. (1976). Working wives: 1940-1960. American Sociological Review, 41, 65-80.

* * *

Discrimination, Education and English: Their Effects on Hispanic Achievement

Franklin J. James and Laura R. Appelbaum

Graduate School of Public Affairs, University of Colorado, Denver

This paper first briefly outlines contemporary indicators of the economic status of Hispanics and how this status has changed in recent years. It also summarizes and assesses the evidence regarding factors shaping this economic status. It highlights critical gaps in our knowledge of how to foster greater achievement among Hispanics.

HISPANIC ECONOMIC STATUS

The conventional wisdom regarding Hispanics as a group is that their economic status lags behind that of Anglos but exceeds that of Blacks. This intermediate status could be viewed as evidence that Hispanics in the U.S. have greater access to economic opportunity than do Blacks. In contrast to the conventional wisdom, the average per capita income of Hispanic households was only 56 percent that of whites in 1983. Black and Hispanic incomes are essentially the same in per capita terms. The median annual earnings of year round full time workers are lower for Hispanics than for Blacks or Anglos. The earnings gap between Black and Hispanic men was very small. However, the median earnings of Hispanic women were around nine percent below those of Black women in 1982-1983. The economic status of Hispanics is lagging behind that of other minority groups in the U.S. The

decade 1970-1980 was a relatively adverse one for Hispanic workers (James & Appelbaum, 1986). The James & Appelbaum study focuses on working age persons with substantial ties to the labor force. The annual earnings of Hispanic men held constant during the decade relative to those of Anglo men. By contrast, the relative earnings of Black men rose markedly.

THE DETERMINANTS OF HISPANIC STATUS

Recent research offers useful insight into the factors influencing the earnings of Hispanics, Blacks and whites. The so-called human capital model provides empirical evidence on how various characteristics of workers shape their productivity, and on the relative importance of potential productivity and discriminatory barriers in determining actual wages or earnings (Mincer 1974). Research using this model has suggested that, among men, the bulk of the wage gaps separating Hispanics and Anglos can be attributed to:

- limited average schooling
- labor market discrimination
- handicaps in the use of English

EDUCATION

Virtually every study has reported that poor education is a principal factor depressing the earnings of Hispanics. In 1980, for example, only 40% of foreign born Hispanics with substantial labor force ties graduated from high school and 9% from college. The comparable figures for Anglo men were 83% and 25% respectively. Native born Hispanic men also were poorly schooled relative to Blacks and Anglos. Hispanics also failed to make a significant dent during the 1970's in the gaps separating their schooling from that of Blacks or, more importantly, Anglos.

One possible explanation for the limited schooling of Hispanics is that the economic payoffs of education could be low for Hispanics. Research by James and Appelbaum suggests that educational payoffs are as high for Hispanics as for Anglos, and that the payoffs during the 1970s. Inadequate incentives do not appear to play a role in explaining the current limited schooling being sought by Hispanics. One recent study used High School and Beyond data to examine the school dropout decisions of high school students between their sophomore and senior years (Fernandez and Hirano-Nakanishi; undated). This study found that the following factors strongly increased the probability that Hispanic students would drop out:

- marriage and having children
- poor grades
- female head of household
- first generation immigrant
- bilingual students, relative to students monolingual in English

The apparent importance of immigrant status and language skills clearly implies that incomplete assimilation into the U.S. and its culture are of importance in producing higher school dropout rates, and, by inference, lower overall schooling.

Labor Market Discrimination

It is readily possible to assess the extent of some types of housing discrimination encountered by minorities through what are termed "audits" or "tests" in which matched pairs of minorities and Anglos respond to advertisements of housing available for rent or sale (HUD, 1979A; HUD, 1979B; James, McCummings and Tynan, 1984; Hansen and James, 1986). Unfortunately, this technique is very difficult to apply to job discrimination. Interestingly, the one such study which has been applied to measure job discrimination found that the English skills of Hispanic applicants shaped their reception by employers (Santos, 1985, p.5).

The primary evidence available to measure job discrimination is disparities in worker earnings which remain unexplained after the most thorough possible effort to account for differences in expected worker productivity. Cordelia Reimers has estimated that discrimination reduced the expected earnings of Mexican origin male workers by 6%; of Puerto Rican males by 18%; and of Central and South American Hispanic males by 37% in the 1970s. In comparison, her analysis suggests that labor market discrimination cut the expected wages of Black males by 14% in 1975. The James and Appelbaum study found no evidence that Hispanics - native or foreign born - made significant progress in overcoming the barriers of discrimination during the 1970s. The data suggest that labor market discrimination and other unmeasured factors reduced the incomes of native born Hispanic males by slightly more than ten percent relative to the earnings of Angle males in both 1970 and 1980. Discrimination undercut the expected earnings of foreign born Hispanic males by around 25 percent in both years. By contrast, Black men encountered much more serious discrimination in 1970 (-36%), but made significant progress in overcoming it. By 1980, labor market discrimination is estimated to have reduced the earnings of Black males by 26 percent, a still high but much lower figure. Available evidence suggests that civil rights agencies such as the U.S. Equal Employment Commission are not as effective in aiding Hispanics as they should be (U.S. EEOC, undated). Evidence on housing discrimination suggests that Hispanics themselves must become more aggressive in seeking the protections offered by civil right statutes (James, McCummings and Tynan, 1984). Much more research is needed to establish how job discrimination against Hispanics occurs, and what public and private strategies can effectively combat it.

Language and English Skills

In recent years, considerable debate has arisen over the reliability of estimates of discrimination like those presented in the previous section. Some recent research has reported limited English proficiency may be so important as to account for virtually all the earnings gap between Anglos and Hispanics left unexplained by educational disparities. The first research to argue this way (McManus, Gould and Welch, 1983) used indicators of language proficiency which relied most heavily on the language used in a person's household, and least heavily on persons' self-assessment of English ability. As McManus pointed out elsewhere, it is at least possible that these measures of language proficiency reflect cultural assimilation

and social class more than language expertise (McManus, 1985). Even the most recent studies have used only subjective indicators of English skills, so that their findings are dubious. These same studies also generally omit direct indicators of a person's likely cultural assimilation into the U.S., almost certainly biasing upwards statistical indicators of the importance on English skills per se. Language may in addition be used by employers as a flag for discriminatory treatment. This conclusion is supported by the experimental research on employer discrimination cited above (Santos, 1985).

CONCLUSIONS

Available evidence offers useful but clearly not conclusive evaluations of possible strategies for improving Hispanic economic performance. Improving the educational achievement of Hispanics is clearly the top priority, but evidence is tantalizingly thin on how to do so. Stronger public and private efforts to combat discrimination in the labor market is also a strategy of potentially great value to Hispanics, as are programs designed to increase the mastering of English among Hispanics. One thing is clear: no one of these strategies is likely to be sufficient alone to significantly upgrade Hispanic status.

* * *

Selection and Assignment in a Large Organization:

Project A

Development and Validation of Army Selection and Classification Measures(1)

Prepared by:

Human Resources Research Organization (HumRRO)
American Institutes for Research (AIR)
Personnel Decisions Research Institute (PDRI)
Army Research Institute (ARI)

Presenter: Douglas Kuhn

Human Resources Research Organization (HumRRO), Alexandria, Virginia

(1) This research was funded by the U.S. Army Research Institute for all Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

INTRODUCTION

The purpose of this paper is to discuss a project entitled: "Improving the selection, classification, and utilization of Army enlisted personnel"--Project A for short. The project is funded through the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and, together with ARI research staff, is being carried out by a consortium of three firms: the Human Resources Research Organization (HumRRO), the American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI). Project A is a nine year project whose overall purpose is to provide the data to design improvements in the selection/classification system for enlisted personnel. The improvements are in the form of developing new classification tests to supplement the Armed Services Vocational Aptitude Battery (ASVAB) and to validate all selection/classification measures against a broad array of job performance criteria. It is, to our knowledge, the largest research and development project ever undertaken in personnel management. The basic requirement is to demonstrate the validity of the ASVAB as a predictor of both training and on-the-job performance.

In reviewing the design needed to meet that requirement, the concept of a larger project began to emerge. With only a moderate amount of additional resources, new predictors in the perceptual, psychomotor, interest, temperament, and biodata domains could be evaluated as well. And a longitudinal research data base could be developed, linking soldiers' performance on a variety of variables from enlistment, through training, first tour assignments, reenlistment decisions, and for some, to their second tour. Finally, those data could be the basis for a new way to allocate personnel, making near-real-time decisions on the best match between characteristics of an individual enlistee or reenlistee and the requirements of available Army military occupational specialties (MOS). Specifically, then, the objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria, the latter to include both Army-wide performance measures based on newly developed rating scales and direct measures of MOS-specific task performance.

- Develop and validate new and improved selection and classification measures.

- Validate intermediate criteria, such as performance in training, as predictors of later criteria, such as job performance ratings, so that more informed reassignment and promotion decisions can be made throughout the individual's tour.

- Determine the relative utility to the Army of different performance levels across MOS.

- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The project is not being conducted as a set of separate tasks that make "inputs" to one another and that are to be "integrated" somehow. Such a view misses the essential unity of the effort; Project A is one project and is organized into five major tasks.

Task 1. Validation

Task 1 has two major components. The first component is to maintain the data base and provide the analytic procedures to determine the degree to which performance in Army jobs is predictable from some combination of new or existing measures. The second component is to conduct the appropriate analyses to determine whether the existing set of predictors, new predictors, or some combination of new and existing predictors has utility over and above the present system. These two components are being accomplished using state-of-the-art technology in personnel selection research and data analytic methods.

Task 2. Developing Predictors of Job Performance

To date, a large proportion of the efforts of the armed services in this area have been concentrated on improving the ASVAB, which is now a well-researched, valid measure of general cognitive abilities. However, many critical Army tasks appear to require psychomotor and perceptual skills for their successful performance. Further, neither biodata nor motivational variables are now comprehensively evaluated. It is in these four non-cognitive domains that the greatest potential for adding valid independent dimensions to current classification instruments is to be found. The objectives of Tasks 2 are to develop a broad array of new and improved selection measures and to administer them to three major validation samples. A critical aspect of this task is the demonstration of the incremental validity added by new predictors.

Task 3. Measurement of School/Training Success

The objective of Task 3 is to derive school and training performance indexes that can be used: (1) as criteria against which to validate the initial predictors, and (2) as predictors of later job performance. Comprehensive job knowledge tests were developed for the sample of MOS investigated and their content and construct validity determined.

Task 4. Assessment of Army-wide Performance

In contrast to performance measures which may be developed for a specific Army MOS, Task 4 will develop measures that can be used across all MOS (i.e., Army-wide). The intent is to develop measures of first- and second-tour job performance against which all Army enlisted personnel may be measured. A major objective for Task 4 is to develop a model of soldier effectiveness that specifies the major dimensions of an individual's contribution to the Army as an organization. Another important objective of Task 4 is to develop measures of utility. It is critical to define the benefits likely to accrue from what will probably be more costly selection/classification procedures.

Task 5. Develop MOS-Specific Performance Measures

The focus of Task 5 is the development of reliable and valid measures of specific job task performance for a selected set of MOS. This task may be thought of as consisting of three major components: job analysis, construction of job performance measures, and construct validation of the new measures. While only a subset of MOS will be analyzed during this project, the Army may in the future wish to develop job performance measures for a larger number of MOS. For this reason, the methods are intended to apply to all Army MOS.

General Outcomes

The Project A Research Plan speaks to the specific operational and scientific outcomes that will flow from the project. They are characterized by the following themes:

- Project A will generate a broader and more complete sample of the predictor space than has ever been used before in a selection investigation. The taxonomy of predictors that is established will stand as a reference point for many years to come.

- Project A will provide the most thorough attempt ever made to develop standardized tests of task performance in skilled jobs. The procedure used will stand as a model.

- Project A will be the most thorough test to date of whether success in training predicts success on the job.

- Project A will provide a state-of-the-art model to illustrate how construct validity can be used to study applied problems in selection/classification and performance assessment.

- Project A will be the first large selection and classification research effort to incorporate utility in the development of operational decision rules.

- Given the broad range of predictors, criteria, and jobs, Project A will be the most comprehensive evaluation ever conducted on questions of differential predictability across jobs, criterion measures, and predictor constructs.

We believe that Project A will make significant contributions to improve Army operational capability and to provide the most satisfactory careers for individual soldiers. Further, we expect that substantial scientific development will result from this effort.

Complete documentation on the analyses and results of the criterion measures field tests is presented in the following four documents:

Davis, R.H., Davis, G.A., Joyner, J.N., & de Vera, M.V. (1985). Development and field tests of job-relevant knowledge tests for selected MOS (Draft). Alexandria, VA: Human Resources Research Organization.

Borman, W.C., & Pulakos, E. (Eds.) (1985). Development and field test of Army-wide rating scales and the rater orientation and training program (Draft). Alexandria, VA: Human Resources Research Organization.

Campbell, C., Campbell, R., Ramsey, M., & Edwards, D. (1985). Development and field test of task-based MOS-specific criterion measures (Draft). Alexandria, VA: Human Resources Research Organization.

Toquam, J., et al. (1985). Development and field test of behaviorally anchored rating scales for nine MOS (Draft). Alexandria, VA: Human Resources Research Organization.

* * *

UNIQUE PUBLIC SECTOR EXPERIENCES: SPECIAL PROBLEMS AND SOLUTIONS (Paper Session)

The Administration of a Sanitation Worker Physical: Challenges and Solutions

Esther K. Juni, New York City Department of Personnel

The task: Administrate a physical abilities test for Sanitation Worker to over 62,000 people, including 3,000 women. The time frame: completion within a year. Duration of the test: 27 minutes per candidate.

The job of Sanitation Worker in New York City is a highly desirable one. The starting salary is $23,104 and after three years automatically rises to $29,619. Retirement is at half pay after 20 years. There are no education or experience requirements. Thus, it was not surprising that over 62,000 people took and successfully completed the first part of the examination, a pass-fail written test. For the test administrator, finding a site large enough to test 62,000 people for 27 minutes each and complete testing within one year presented a problem. The site finally chosen was an unused aircraft hangar, known as the Blue Nose. To make the Hangar accessible to the candidates, the City provided free shuttle van service between the nearest subway stop and the Hangar.

Once the site had been chosen, test equipment could now be designed. The test was modeled after the duties of a collection worker. Candidates, like collection workers, would begin with a pile of garbage, pick up and deposit the bags in a simulated hopper, wait (while the bags were cleared) and then walk to the next pile of garbage. There would be eighteen such piles of garbage. Candidates would continue loading and waiting until they had loaded the last pile of garbage bags. This process was to last 27 minutes, with the total weight of the garbage to be lifted in that time set at 2975 pounds. The weight of the individual garbage bags would range from 8 to 65 pounds. One of the most difficult problems we faced was simulating garbage bags and their contents. We finally settled on United States Postal Service air-mail bags and leather scrap for the contents. The bags were filled to the prescribed weights with the scrap leather. Another problem was the design of the receptacle which would allow bags to be thrown into it by one candidate, yet would not require that the bags be placed back into their original starting position for the next candidate. It was agreed that the height of the receptacle into which the bags would be thrown would be the actual height of a garbage truck, 38 inches. Finally, we hit upon a simple solution. A U-shaped band of metal 38 inches high with wheels attached to the bottom was designed. This could be locked into a metal back stop on either side. Thus, candidate one would simply throw the bags over the U-shaped band of metal onto the floor then go on to the next group of bags. The examination monitor would then just turn the U-shaped band of metal around and attach it to the backstop on the other side. The next candidate would pick-up the bags from where they were thrown by the previous candidate and lift them over the metal band which was now on the opposite side. If the previous candidate failed to lift a bag, the monitor was required to move the bag to the side where the other bags in that group had been thrown. The backstops were the height of the inside of the garbage truck - about eight feet.

The major testing issue that remained was the timing of the test. Sanitation Workers in New York City performing collection duties are required to work at a steady pace. They are expected to finish their route during a tour of duty. Thus, a timing sequence had to be devised which would require candidates to work at a steady pace and would also include the time for clearing the truck, when a sanitation worker simply waits for the garbage to clear, and then walks to the next pile of bags.

Made to order timers which consisted of a box with large digits containing a red and green light were purchased. The red light counted down for thirty seconds, (from 30 to 0) while the green light counted down for sixty seconds. Timers were placed on top of the middle backstop in every set of three. Timers automatically went from the green light cycle (60 seconds) to the red light cycle (30 seconds) three times, coinciding with the three piles of bags at each house or station. When the red light came on for the third time, it automatically activated the red light in the timer at the next station. When it completed that 30 seconds countdown, that timer went blank and timing was continued by the next timer. This continued for all six interconnected timers. By the time the last timer turned red, 27 minutes had elapsed.

Candidates were simply instructed to obey the lights. When the green light was on, they were to load bags. When the red light came on, they were to stop loading, walk to the next group of bags and wait until the green light came on. (This simulated waiting until the garbage in the truck had cleared and then walking to the next group of bags.) They then resumed work. Candidates began the test by pushing a Start button which activated the clocks. This caused a red light to go on and allowed the candidate to walk to the first group of bags before the first green signal went on. Four candidates could be on each course simultaneously.

The administration of the test was the final hurdle to be overcome. Candidates were called in social security number order, every half hour from 8 a.m. to 5:30 p.m. After signing in and being given waiver forms to fill out, candidates were led to the video room and shown an orientation film which informed them of the rules. Candidates had previously received a handout describing the test. After the film, candidates were led to the test course and given numbered bibs in one of four colors. Each color represented one of the four parallel tracks. They were also provided gloves in one of three sizes, to wear during the test. They were seated, fingerprinted and called in bib order to take the test.

Women were not called separately, but were interspersed with the men in the order of their social security number. The only exception to this call were women who participated in a special training program conducted by the Center for Women in Government, a non-profit agency. These women were called at the completion of their training program. Separate facilities (lockers and bathrooms) were provided for each sex.

An examiner was assigned to each candidate for the duration of the test. The examiner watched to make sure the candidate followed instructions, watched the timer and stopped work when the red light came on. All uncollected bags - i.e., bags that the candidate failed to load within the required time - were noted on the examiner's sheet. At the end of the test, the candidate was rated by tallying the number of bags listed as uncollected on the examiner's sheet, and was informed of his/her score. The scoring system, as shown in the film was as follows: All bags collected within the prescribed time - 100% - Band 1; from 1-15 bags uncollected weighing no more than 300 pounds uncollected - failure.

The weights of each bag were imprinted on the leather strap used, in conjunction with wire, to tie the bag together. The wire binding and leather imprinting were done by the scrap leather dealer from whom we purchased the leather. The examiners used on this test were college graduates with training in Physical Education.

Since the test ran six days per week (Monday-Saturday) and each day was a twelve hour day, two shifts of personnel (both examiners and monitors) were used each working three days per week.

Safety was another administrative concern. A special medical room was built into the hangar for emergencies. It was staffed full-time by two emergency medical technicians.

To sum it all up, it is possible to conduct a twenty-seven minute physical test for each of 62,000 candidates and complete testing within one year.

* * *

Planning and Conducting an Assessment Center in a Strong Union Environment

Donald G. Bergeson, City of Miami, Florida
Lawrence R. O'leary, Ph.D., O'leary, Brokaw & Associates, Clayton, Missouri
C. Dan Fabyan, Deputy Fire Chief (Retired), City of Miami, Florida

The selection of Line Officers for today's Fire Service is a major concern of Fire Service Administrators in the 1980's and will continue to be a concern in the future. Mandates for equal opportunity and affirmative action, along with concerns for better managers have made it necessary for Fire Service Administrators to scrutinize their traditional methods of testing which usually consist of paper and pencil tests along with credit for seniority. The City of Miami replaced its paper and pencil job knowledge test for chief fire officers with the assessment center method. This method of promotional testing would be a radical change from traditional testing methods used in the past by the City of Miami.

The results of the joint management-union report were distributed to key personnel for review and were part of efforts to create a climate for change. For over a year, informal discussions were held with potential candidates, key union officials and city administrators in an attempt to inform everyone concerned of the beneficial aspects of assessment center testing. These efforts paid off when in the fall of 1984 the union agreed to assessment center testing for the chief fire officer exam scheduled for June of 1985. The Department of Human Resources, which is responsible for all promotional testing, agreed to play a major role in coordinating the project. A description was made of the procedures followed to secure funding, select a consultant, and implement the project under a contract.

A systematic job analysis of the position included: 1) An organizational chart of the overall department; 2) Existing job descriptions of the Chief fire officers position, as well as those of Captain and Lieutenant; and 3) Interviews with more than 50% of the incumbent chief fire officers from all shifts and Districts.

The job analysis procedure was a combination of a method known as the Retrospective Critical Incident Approach and some on-job observations. These techniques included a verification phase, in which the initial findings were shared with a large population of job experts in order to involve the input of as many knowledgeable people as possible. A standardized interview format was used; interviews varied from one to four hours in length. The longer sessions included observations of the incumbent during "emergency runs." In addition, all of the people who directly supervise chief fire officers within the Miami Fire Department were also interviewed.

Another perspective, besides that of the incumbent, is the supervisor's view. Consequently, all three Division Chiefs who directly supervise chief fire officers in the Miami Fire Department were interviewed. The specific emphasis in these interviews was to review critical incidents which reflect particularly high performance as a Chief fire officer, or conversely, particularly low performance. In addition, the major tasks required of this position were requested. These interviews took between one and one and a half hours as a rule. The above interview schedule resulted in the consultant directly interviewing and observing 9 of the 16 Chief fire officers and all of the Division Chiefs who supervise that position. This was considered more than sufficient to adequately identify the requirements for the job.

Based on the material obtained in these interviews, the consultant pulled together two lists:

1. A list of major competencies required to perform the duties of chief fire officer.

2. A list of the major tasks required of the chief fire officer.

These two documents were then presented to all of the chief fire officers for verification, as to whether they were important for the job and whether they described the major responsibilities of the chief fire officer position. Consequently, a list of competencies measurable by the assessment center method was circulated among all the Captains for their input. They were asked to identify the 15 most important competencies listed in the document as they related to the position of chief fire officer for the Miami Fire Department. They were then to spread 100 points across those 15 competencies to indicate the relative importance of the 15 competencies. This input, in conjunction with the interviews with the chief fire officers and the Division Chiefs, resulted in the selection of the final 12 competencies to be measured in the assessment center.

Based on the job analysis, and more specifically, the competencies identified as important for the job, five exercises which tapped these competencies were selected. The following is a description of the five exercises.

1. Analysis Exercise

The decision to measure technical skills in the assessment center had been made before the consultant had begun his work. The only decision left was whether one or two exercises would be included to measure these technical skills. The consultant's strong belief was that the assessment center was a measure primarily of supervisory and management skills, but that it could give a general indication of the level of technical expertise.

Consequently, one exercise, the analysis exercise, was identified as a point which would give the candidate an opportunity to display the level of technical competence. The analysis exercise was developed jointly by the consultant and subject matter team, composed of Chief of Operations, Chief of Firefighting and Chief of Training.

2. In-Basket

The In-Basket was tailor-made by the consultant based on (1) his review of the chief fire officers actual In-Baskets, and (2) a review of all forms used by the department.

3. Coaching/Counseling

The coaching and counseling exercise was also tailored to the Miami Fire Department and based on information about typical types of counseling situations in which a chief fire officer and a captain might interact.

4. Group Discussion

The group discussion was also based on information gained in the job analysis. Issues which had no obvious right or wrong answers, but were considered pertinent to the Miami Department of Fire and Rescue, were selected.

Each of these exercises were reviewed by three members of the subject matter expert committee and judged to be at the same level of difficulty as that found in the job of chief fire officer.

5. Background Interview

The background interview rationale was the fact that motivation and other dimensions could be measured by a background interview. In point of fact, motivation could be measured in no other exercise. Because motivation was considered an important competency for the job and because the only other selection tool that would be involved in formulating a person's rank on the list would be his seniority score, the consultant recommended the inclusion of the background interview as an exercise.

Approximately one month before the assessment center, a packet of material, including a number of articles about assessment centers and a bibliography for further reading on assessment centers were provided each candidate. In addition, an orientation session for all interested candidates was conducted at the Miami Fire Academy on April 29th.

A video tape showing actual assessment center exercises was shown. In addition, the consultant explained the basic component of an assessment center, and how a candidate could do his best. Finally, the candidates were encouraged to ask any questions about the assessment center and the entire promotion process. This session lasted for approximately two hours.

The assessors were selected by Chief of Operations (10 assessors) and the consultant (1 assessor). All assessors were executives in large fire departments at a level higher than that of chief fire officer.

Four days of training (May 30-June 3, 1985) were devoted entirely to training the assessors which was conducted by the consultant. The training included reviewing each of the five exercises, learning about the specific mechanics of an assessment center, practicing and more practicing of observation and note taking, scoring and assessor discussion.

The eleven assessors were generally from the level of assistant chief in their respective departments. The group included a number of protected class representatives and came from the following organizations: the Los Angeles Fire Department; Jacksonville Fire Department; St. Louis Fire Department; Dade County Fire Department; Hallandale Fire Department; Albuquerque Fire Department; West Palm Beach County Fire Department; Fort Worth Fire Department; Arlington County Fire Department; St. Petersburg Fire Department; and the Phoenix Fire Department.

The decision was made to use professional actors rather than actual fire-fighters or fire captains in some exercises, in order to maintain confidentiality and because of a greater acting capacity present in professional actors. In point of fact, candidates and assessors alike commented on the high level of credibility in the performance of the role players.

Another step to lend credibility was to dress the role players in fire captains uniforms. Two days were spent by the consultant and the principal author in practicing and critiquing the portrayal of roles described in the coaching and counseling exercise.

The results of assessment center were provided in two forms, numeric and narrative data. The numeric data was comprised of a weighted score for each of the twelve competencies. These summed weighted scores constituted the candidate's final assessment center score. The scores ranged from 52.38 to 103.00, with a distribution mean of 83.0753, and a standard deviation of 16.6804.

One interesting observation was that the seniority and assessment center performance were virtually unrelated. This suggests that the assessment center is measuring something completely different from "longevity on the job." Another conclusion, which this data supports, is that beyond the minimum qualification of two years as captain, effective performance in the assessment center is not more likely if you are an older, more experienced captain, or younger and less experienced.

The final promotion list was based on 80% final assessment center score and 20% seniority. Feedback sessions with the candidates were conducted in which strengths and weaknesses according to competencies were discussed by the principal author. The feedback sessions were reviewed as educational by most candidates. In fact, most of the candidates agreed that the strengths and weaknesses pointed out by the assessors were accurate.

Before the results were made public, the chief of operations conducted a group exercise entitled "nominal group process," in which all candidates had an opportunity to discuss the negative and positive aspects of assessment center testing for chief fire officers. The results showed that all were in favor of the new process and felt it was a fair and meaningful way to test for the skills and abilities needed for today's fire service managers.

When the critical comments about the assessment center process were compiled, the most frequently mentioned criticism was, "there should be more than one assessor observing each candidate" and a comment critical of the limited overlap between the reading list and the analysis exercise.

In conclusion, the City of Miami took a great deal of time and expense to develop a valid selection system. The end result was viewed by all concerned as worth it. An additional indication of the validity of this conclusion is the Department's intention to use the assessment center to create the next promotional list for chief fire officer.

* * *

Making Merit Systems Work - An Unconventional Approach

Geoffrey Rothman

San Francisco Civil Service Commission, California

The San Francisco Civil Service system has both its design and specific operating procedures codified in a charter which is amendable only through popular vote or judicial ruling. It is administered by a five person lay

commission appointed by the Mayor for staggered terms of six years each, and who are removable only through an impeachment process. With the exception of labor relations, most personnel matters are subject to the authority of the commission. The delegation of personnel authority to departments is particularly restricted in the areas of selection, classification, and compensation. In the area of selection, from eligible lists, for example, San Francisco uses a rule of three names, which substantially limits departmental managers in hiring decisions.

Additionally, sections of the Charter spell out precisely detailed procedures, for example, in the area of promotional testing in the Police and Fire departments, including designating the types of tests to be used, protest and review procedures, and detailing point awards for seniority, merit and education.

There are several consequences of this extremely rigid, rule bound, system. The first is that creative and innovative approaches have, over time, developed to circumvent the most extreme constraints of the system, allowing for flexibility and adaptability to day-to-day needs. The second consequence is the regular collisions that occur between what sometimes appear to be two parallel systems, the formal one and the informal one. A third outcome is the effect on the nature of management. To illustrate the effects that these factors generate with regard to the selection and retention of employees, I will discuss the evolution of the provisional employee program.

San Francisco city government's history is, prior to the 1930's, so notorious and colorfully tarnished that the regulatory nature of the Charter, particularly in the personnel function, is not surprising. However, even a group of reform minded civic leaders did not preclude the possibility that some few persons might need to be employed without competitive tests. In fact, they created one exceptional employment category called 'Non-Civil Service' to cover short-term and temporary personnel needs that could not be efficiently serviced by the examination program. However, this one exception to merit system employment was limited to only those situations where no exam lists were available, and allowed employment which was restricted to a maximum of ninety working days in a year. There were no significant change to this personnel system for some time.

Beginning with World War II a dramatic shift occurred in the labor market. Because the Charter framers had not contemplated a large scale migration by a sizeable number of municipal employees for more than a ninety-day period, the City reacted by creating a provisional employee program, to replace the vacant city jobs known as Limited Tenure. This approach allowed a great percentage of city employees to do their patriotic duty with the assured availability of their old jobs upon their return. In the meantime, their jobs could be filled with temporary workers who would not be required to take and pass Civil Service exams and who would not gain any right or preference to their positions. The enabling ordinance restricted such a departure from the merit system to times of war and the national draft.

By the 1960's, other uses of the limited tenure programs began to emerge. The city realized that there was no need to find benefits for temporary workers and as such the limited tenure program represented an excellent cost savings vehicle.

The limited tenure population increased yearly with the decline in examination productivity, the greater number of classifications requiring tests and the expanding total workforce. Another contributive cause was the hiring control that the limited tenure program delegated away from the traditional testing program and gave to the departments. In effect, the limited tenure program allowed departmental managers to hire qualified employees, but without regard to examination lists, or other merit system controls and procedures. This approach also met the needs of many employee training efforts and affirmative action programs.

Beginning in 1981 several new factors began to emerge which signalled the beginning of the end for the limited tenure program. First, the Civil Service examination program was reorganized for the primary purposes of increased accountability and productivity. This action was the single most significant factor in contributing to the end of limited tenure. This action had two direct consequences. First, with an increasing number of tests being given, departments were threatened with severe disruption because of the loss of many temporary employees who did not perform well in highly competitive testing.

With the advent of agency shop, organizations were brought into the picture. As these unions accepted dues from the limited tenure personnel, they were obligated to represent their interests. At this point the unions were caught in a number of dilemmas including a divided membership, lack of leverage, no tangible solutions to recommend, and a substantial lack of success in pursuing administrative and judicial relief for displaced limited tenure employees.

The problem of possible displacement of a large percentage of limited tenure minority employees was becoming critical. A disproportionate number of minority employees were being displaced, frequently by other outside minority candidates in the more rapid job-related testing effort.

Eventually, in early February 1983, a document titled the Temporary Employees Letter of Agreement (LOA) emerged. The LOA contained three key ingredients. First, an accelerated testing program was to commence immediately, with nearly two hundred tests to be completed in about five months. The majority of these examinations would be in the form of training and experience ratings, using an unassembled testing procedure. Secondly, Civil Service was to receive funding sufficient to bring its examination program up-to-date and maintain that level indefinitely. Third, a two million dollar fund was created to fund positions for longterm temporary employees who were displaced as a result of this examination program. The agreement was signed by the mayor on February 18, 1983. All examinations had to complated by August 30, 1983.

Three distinct problems were presented to the examination unit. The first was one of test construction. Specifically, what kind of job analysis would be utilized and how would it translate into minimum qualifications and competitive test components. The second problem was logistical, demanding a highly efficient use of all available personnel and resources, including the maximum utilization of a dozen totally new and untrained examiners, and an orderly job announcement, application, testing and appeal process. Third, all of these factors had to be integrated to achieve the objective of transitioning a substantial number of limited tenure employees.

The first step in the solution involved selecting an experienced examiner to supervise this kind of an effort. In making that selection I utilized a job/person matching system derived from motivational theory. Specifically, I charted out the motivational profile of the supervising position on three scales including achievement, affiliation and power and supplemented that profile with other vital factors. The supervisor was in turn given about three weeks to develop a program scheme including comprehensive details and procedures. At the same time the supervisor elected to hire a totally new staff of examiners to be the primary program team. Only about two thirds of these examinations were handled by the special unit, designated the ATP unit. The balance of examinations were farmed out to other existing examination units. This approach allowed for a more effective utilization of the total staff, while concentrating program coordination and primary program responsibility in only one unit. The new staff was composed, principally, of recent college graduates.

The next task was to decide on an examination plan. Unassembled examinations were commonly utilized for semi-skilled entry level classes such as Inventory Clerk, Homemaker, Janitor, and Cashier. The assembled examination group included senior or supervisory level classifications. The Health care occupations were generally included in the assembled test group regardless of level, and were the only notable exception to this assignment pattern.

An abbreviated job analysis procedure was used with the job activity and KASO (Knowledge, Ability, Skill and other characteristics) or job element data being derived from a combination of the prior completed job analyses, the classification specification, and one-on-one review and verification of resulting information with subject matter experts. A generic test plan was applied to each unassembled examination utilizing major job activity statements as a basis for rated supplemental application and using the KASOs to form the basis for the minimum qualifications. In the case of the rated applications, all major job activities were listed and rateable.

Each candidate was asked to indicate how much experience they had performing each activity. All activities were assigned equal value for purposes of final summary ratings. There were several rateable experience levels in six months increments from no experience to more than forty-eight months of experience. To verify experience claims each applicant was required to submit an employer's letter detailing length of employment, job title, typical duties, etc. Any claim that was unverified was denied and no points were awarded. For activity statements that were difficult to rate

due to imprecise verifications subject matter experts were consulted to award final scores. The same basic system was used in developing minimum qualifications. The KASOs were converted into specific training and experience requirements by examiners and verified by subject matter experts. Only claims which had accompanying employers or training verifications were acceptable. Because this system provided no opportunity for evaluation of oral communications, an additional condition was established whereby any candidate could be refused consideration under the Rule of Three certification if they did not possess adequate English language fluency as judged by the appointing authorities.

The second major challenge of the program was the issue of application and testing logistics. Specifically, how does an agency announce nearly two hundred examinations, handle applications, and applicants in an expeditious manner and produce a great volume of eligible lists, while anticipating a huge onslaught of protests and appeals. The initial key to meeting this challenge was ensuring the close cooperation and support of the labor unions who had signed the agreement.

To ensure the simplest, most expedient and most economical approach to announcing these examinations, most of the examinations were listed on one twenty-page examination bulletin. Additionally, because of printing costs, and logistical difficulties, copies were publicly posted but were not physically distributed to candidates. Information about each examination class was briefly presented along with descriptions of the application procedure. In the case of unassembled examinations a special application was required for each classification. To organize the application filing process a procedure was created whereby, over a three-week period, pre-numbered applications for each class were distributed on an in-person basis on one specific day. Likewise, approximately three to four weeks after the application pick-up date there was an application filing date. Applications could be filed in person or by mail if postmarked on the application acceptance date. This somewhat complex inflexible system resulted in the distribution of more than thirty-thousand applications, with the return of less than half.

The most controversial element of the testing program, and subsequently the only significant issue to reach the Civil Service Commission on appeal, involved the candidate reduction technique known as 'series' testing. Specifically, it was decided, as a rule of thumb to test no more than eight candidates for each then existing vacancy. Eventually, this matter came before the Civil Service Commission, which reaffirmed the concept, but modified the applicant to vacancy ratio to ten to one.

The evaluation of this effort falls into many categories. First, and foremost the program was successful in meeting its schedule and most of its major goals. Although, no exact count has been conducted it appeared that better than half of the longterm limited tenure employees were transitioned to permanent status. Conversely, a significant number of limited tenure employees were displaced. They were, however, covered by the insurance aspect of the LOA, and were kept in municipal employment until December 1984. The two million dollar fund was fully expended. Based on the

measure of the third program objective, transitioning limited tenure employees, and curtailing the use of the limited tenure status, the program was highly successful.

Unfortunately, the Accelerated Testing Program cannot be measured against any one criterion. In addition to celebrating the apparently successful ending, we must measure the program against several other factors.
First, in practical terms, what did the program achieve? As noted earlier, the program did operate successfully to eliminate most longterm limited tenure appointments. It did represent an effective combination of labor, management, and political interests in an effort which brought some advantage to each principal participant. However, by curtailing the limited tenure program, numerous deficiencies in the current Civil Service legal environment became more obvious and more problematic. The major weaknesses continued to be the unduly complex and delaying process of examination appeals and the several restrictive effects of the Rule of Three.

The Rule of Three and administration of eligible lists are other major dilemmas in the Civil Service selection structure. Although, this restrictive employment procedure is well intended to ensure that only the most meritorious are employed, the practical effect is to focus most controversies on the testing program. This is compounded further by the long duration of the examination lists, a two-year minimum, and the inability to utilize subsequent lists until the most senior lists have been exhausted. The limited tenure program easily circumvented this problem by effectively allowing for minimally restricted hiring delegated to line management.

As noted at the beginning of this paper, the result of focusing on these types of problems in part led to a reform effort to modernize the Civil Service system. This effort was strongly opposed by labor and the majority of their political allies. The major objective of the opponents of reform appeared to be the introduction of collective bargaining as a prerequisite to Civil Service reform.

One of the other less obvious, but predictable, results of this effort has been a substantially increased cost of government. This cost increase has come from two sources including higher costs for Civil Service operations, and a much higher citywide personnel cost.

Since 1983 there has been a gradual restoration of a full merit system. Even the Charter reform effort provides evidence of a revitalized interest in a more responsive and more efficient personnel system, operating in accordance with its legal mandate.

In conclusion, the Accelerated Training Program can be viewed as a real life example of the durability of merit systems in ingenuity of merit system managers in the public sector. The program offers proof that merit system principles can be successfully adapted to solve a wide range of virtually insurmountable personnel problems as well as deliver well qualified eligibles on a routine basis. This case study reveals the underlying strength of merit system concepts to adapt to the challenges and changing environments of contemporary public sector organizations.

PERFORMANCE APPRAISAL: DIRECT APPLICATIONS FOR SELECTION (Paper Session)

Behaviorally Anchored Performance Evaluation Development, Implementation and Results

Foster Dieckhoff, City of Kansas City, Missouri

The "behavioral anchored" approach to performance evaluation is, perhaps the most "face valid" instrument to date. Variables such as the degree to which the "dimensions" addressed are unique to a given position and the number of specificity to benchmarked behaviors must be addressed. In an attempt to address this problem, several classification schemes were investigated. The one finally adopted was a modification of the Occupational Groups used in our classification system. The groups were originally designed by PAS. The resulting groups were named: Clerical, Fiscal, and Administrative Support; Public Safety (except Fire and Police); Technical, Skilled Trades, Recreation and Related Support; and Professional Technical and Staff Support. (The job title for each of the above groups were displayed in slides). The exempt management classifications with "open range" salary structures, were and remain under the MBO system. It should also be added that a special form was developed for Fire Supression classes. The Police Department, which is under the Board of Police Commissioners, is not in the City Merit System.

The next step was to write job dimensions and behavioral benchmarks which were sufficiently specific to be anchored to actual job behavior, but at the same time, be sufficiently abstract to be relevant to more than one job. Interviews with incumbent and supervisory personnel brought to light some "generic" dimensions. The two most common of these that supervisors felt unable to address on the traits-based form were: 1) the employee's tendency to act on those tasks most important to the overall mission of the work unit; and 2) the employee's constructive use of work time. The first we called "Establishing Priorities" the second "Time Utilization." The final product for the Professional Technical, and Staff Support classes is shown below. Since this was the last form we designed, it also has the benefit of previous experience and is probably the better form.

Professional Technical and Staff Support Classes

This portion of the manual has been designed to assist you, the supervisor, in evaluating employees in those classes listed on the preceding pages. The following is a list of job dimensions upon which employees in the listed classifications may be rated:

Mandatory
- Competence in Designated Specialty
- Dependability
- Establishing Priorities
- Time Utilization

Optional
- Oral Communication
- Technical Equipment Care and Maintenance
- Written Communication

Supervisory Only
- Administration of Personnel Policy
- Delegation
- Priorities
- Training

COMPETENCE IN DESIGNATED SPECIALITY — Mandatory

This dimension addresses the demonstrated competence in a designated speciality as determined by the observed quality and/or quantity of the expected work product. Included here are the timelines, accuracy, and efficiency with which assignments are completed. Such characteristics as attention to detail, problem solving, technical competence, and interpersonal skill may contribute to the rating in this dimension.

(a) Optimal: This employee can be depended upon to consistently produce a quality work product within an appropriate time frame. Projects or assignments utilizing specialized skill and/or equipment seldom, if ever, require corrections or additions. Counseling is generally not needed.

(b) Better: Use this rating if you see the performance in this area as better than satisfactory but not optimal.

(c) Satisfactory: This employee normally produces an acceptable work product within a reasonable time frame. Projects or assignments utilizing specialized skill and/or equipment are normally acceptable with minor corrections or additions. Counseling, if needed, results in long-term improvement.

(d) Marginal: Use this rating if you see the performance in this area as less than satisfactory but not unsatisfactory. Note that sustained marginal performance is unsatisfactory.

(e) Unsatisfactory: This employee's production in a designated specialty is unsatisfactory because of an established pattern of one or more of the following: projects or assignments utilizing specialized skill and/or equipment do not meet professional and/or departmental standards; projects or assignments are not completed within an acceptable time frame. Counseling has resulted in little or no improvement.

This particular form is used for technical fields usually associated with degree requirements or other specialized training. The inclusive dimension "Competence in Designated Specialty" attends to major differences among the jobs listed and the language of the dimension and the benchmarks is still specific enough to conjure up work behavior. Similar benchmarks were developed for the other job dimensions.

In reference to the construction of the benchmarks, notice that all dimensions have 5 benchmarks. The 7 or 9 benchmark option scale is basically a trick to get variance in the system. I think training is a better way to get variance--but more of that later. We essentially want the scale to allow, and encourage, the supervisor to accurately assess the employee's performance.

Keeping these facts in mind, we set out to design a scale that provided clear distinctions among the ratings without getting bogged down in the nuances of language. The entire scheme is centered around "satisfactory" performance. "Better" performance is simply that - better than "satisfactory", and "marginal" performance is not satisfactory but not sufficiently sustained to warrant "unsatisfactory." "Optimal" is consistent "better" performance, and "unsatisfactory" is sustained "marginal" performance. The benchmarks for each dimension follow this same logic.

Some of the elements we found helpful to include in the training were to ask the supervisors to compute the dollar value of the human resources they are managing. An informal survey showed that approximately 1 in 21 had any idea of the amount. This of course gets their attention--most were quite surprised at the figure. The appropriate use of a performance evaluation system becomes more than "paperwork" when 60 to 80 thousand dollars annually get into the picture. Another item we try to emphasize is the fact that the legal system is increasingly viewing jobs as property, and, taking away a person's job is little different than taking away personal property such as a car or television. Finally, we emphasize the fact that most employees have strengths and weaknesses, and we as supervisors are obligated to inform the employee of both. The overall rating need not be some type of mathematical average of ratings given on each dimension; rather it is a reflection of overall performance.

I mention earlier that there is evidence that the individual dimension ratings given on the behaviorally anchored forms displayed more variance than those given on the traits-based forms. To substantiate this claim, we computed a 2x2 chi-square of "satisfactory" vs. other than "satisfactory" for several classes on Mandatory dimensions and Supervisory dimensions. The results are both significant at the .95 level. While this provides some evidence of increased variance, which is good news, it turns out that the instrument is still probably not acceptable for use as criterion in traditional empirical validation study.

A rather interesting result showed up when Dr. Jacobsen was plotting a scattergram of score vs. rating on some early data from a new clerical test and a new rating form. He found that those scoring below 54 (out of a possible 70) were substantially less likely to be rated above satisfactory than those scoring 54 or greater. Unfortunately, all of those scoring 54 or better were not rated above satisfactory; hence the actual correlation between overall rating and test score (r=.1158) is not significant.

However, those scoring 54 or above have a substantially better probability of being rated better than satisfactory than do those scoring below 54. Additional investigation showed over 90% of the less-than-satisfactory overall ratings were attendance related. Unfortunately, the examination was not designed to measure attendance, nor does attendance seem to be influenced by test score in the same way as the overall rating. (Note: The statistical tables can be furnished from the author).

To conclude, we have reviewed the construction of a behaviorally anchored performance evaluation system applicable to several job classes within a job "family." The development of appropriate Mandatory and Optional dimensions, as well as their corresponding benchmarks, were outlined. Several elements of the training program used to initiate the Behaviorally Anchored program were also discussed.

Finally, statistical evidence was provided which demonstrated that employee ratings tended to be "other than" Satisfactory (on Mandatory and Supervisory dimensions) at a higher rate on the Behaviorally Anchored form than on the Traits Based form. It was also shown (for Clerical classes) that those who score well (on a content valid exam) have a significantly higher probability of having a "Better" or "Optimal" employee as measured by the Overall rating. It was also noted that test score had no such relationship with attendance—the most common reason for less-than-satisfactory performance. The obvious lesson here is that the criteria used to substantiate empirical validity should ideally measure exactly what the test measures, nothing more and nothing less. In an interesting way this observation lends some credibility to content validity.

* * *

Implementation and Evaluation of A System Using Departmental Ratings For Promotional Decisions

Rodney B. Warrenfeltz, Colorado Department of Highways, Denver, Colorado

During the period between April 1, 1985 and December 1, 1985, the Colorado Department of Highways (Personnel Branch) implemented a new exam process for the positions of Highway Foreman and Senior Highway Foreman. This new exam process represented an attempt to ameliorate a number of problems concerning the promotion of employees within these classes.

The history of exams used in maintenance type positions at the CDOH demonstrates a clear need for a more systematic approach to the selection of promotion candidates. In response to this need, a research project was started to develop and implement a new exam process for maintenance personnel.

The development phase began by designing a flow chart that characterized the exam process from beginning to end. The flow chart was extremely important from the standpoint that it took into account the dynamic nature of maintenance type positions and provided a systematic method for updating exam materials as significant position changes occurred. The flowchart, located in Appendix A, outlines an exam development phase and an exam administration phase.

As can be seen in Appendix A, the exam development phase begins with identification of subject matter experts (SMEs) that are used to obtain job analysis information. The job analysis information is used in updating the Promotion Performance Appraisal (PPA) which will be described in detail in a later section. The information is also used in developing written/oral examinations. The written/oral examinations served two purposes in this process. First, data from these examinations have been used in the PPA validation process in the form of criteria. Second, in the examinations for some positions, the written/oral questions have been used to form a second level of evaluation.

In addition to the development phase, Appendix A also illustrates the administration phase. This phase includes screening of applications, completion of the PPA, and completion of a written/oral examination if the position called for it. From the data gathered during this phase, scores are determined for each applicant and a promotional list is established.

It is also important to point out that SMEs used in the administration phase also have an opportunity to provide update information in the exam process. The information is collected after the administration phase is complete and is used in future updates of the exam process. This information adds significantly to the dynamic nature of the exam process by allowing for an almost continuous flow of update information.

The primary component of the exam process is the PPA. The PPA was developed to assess an applicant's job performance in a promotional context. In other words, in rating an applicant on the PPA, a rater is asked to view the applicant in the context of the new position and to rate on the basis of how well the applicant would perform if promoted.

The PPA form contains two sections. The "Performance Factors" is used for recording job relevant behaviors on the basis of behaviorally defined performance factors. The section begins with a brief set of instructions and is followed by a series of factors that include definitions and behavioral examples. The rater would first be required to document performance behaviors relating to the various performance factors. The behaviors would be examples of an applicant's performance which would provide an indication of ability to perform at the Senior Highway Foremen level. The difficulty of this task is lessened to some degree by providing the rater with examples of relevant behaviors under each factor.

When the rater completes the documentation of behaviors, section B "Performance Rating" can be completed. Section B includes a brief set of instructions, a rating scale and a place to rate each factor. The instructions ask the rater to rate an applicant on the basis of how well the individual would perform if promoted. The rating scale is essentially a five point Likert scale (5=Outstanding to 1=Below Average) that varies along the dimension of an applicant's ability to step into the new job and perform on the factor. This scale is similar to one used by Maher (1985). The rater, after assigning a rating to a factor, multiples the rating times a factor weight. The weights are derived from the SMEs during the updating of the PPA. An applicant's final score is determined by summing across the weighted factor scores to obtain a total score.

Rating Errors

In all of the recent attempts to use rating scales with maintenance type positions, the ratings were found to be replete with rating errors. The type of error most often encountered has been a leniency error or a tendency on the part of the rater to inflate the score of the ratee (Nunnally, 1978). This has the effect of compressing the variance between ratees and reduces the ability to discriminate between good and poor performers. Furthermore, if rating errors of this type occur in an inconsistent fashion, it may result in the over-representation of particular groups in a test or on a promotional list.

To offset this type of error, two procedures were designed into this exam process. The first procedure involved the use of the Ratings Distribution Check Form by the reviewers in a particular exam. The reviewers in all exams using this process are the supervisors of the raters. The reviewers, in addition to applying the Ratings Distribution Check Form, are responsible for checking the ratings in their area for overall accuracy and completeness. The purpose of the Rating Distribution Check Form was to provide reviewers with a forced distribution designed to guard against rating errors.

The second procedure used to guard against rating errors was a post hoc intervention that allowed for the rescaling of scores based on the distribution of ratings obtained in a particular exam. In general, the rescaling procedure involves the determination of a grand mean from all the ratings for a particular exam and adjusting the scores, by a particular unit (e.g., scores within departments, districts, or other organization units), to the grand mean. This has the effect of placing all of the organizational units on the same rating scale with a midpoint equal to the grand mean. See Appendix B for a graphic representation of the rescaling procedure.

It is important to point out two assumptions that are necessary if the rescaling procedure is to be applied. First, there is an assumption that the applicant pools, as a whole are equally productive across organizational units. At the CDOH, this assumption was fairly safe since there was no reason to assume that one district within the state was any more or less productive than another district. In addition, recent productivity data

obtained from SMEs tended to support this assumption. The second assumption needed to apply the rescaling procedure is that productivity is normally distributed among applicants within organizational units. While the procedure is relatively robust in regard to this assumption and the Ratings Distribution Check form helps to insure a normal distribution of ratings, this assumption should be checked if sample sizes permit.

Item Bank

Finally, some mention should be made of the fact that many of the exams associated with maintenance type positions often require the use of a second screening device (e.g., oral or written essay exams). Since the applicant pools for these exams could be reduced in size by setting a cutoff score on the PPA and selecting only the top applicants for the second screening, many of the problems outlined in the introduction could be averted. However, to further reduce the problems associated with oral or written exams, a computer based item bank was started. Since the exam process includes systematic updating procedures, the item bank is continually updated with questions for applicants by the SMEs. This is done in such a way as to reduce the need to continually contact SMEs to obtain question information.

The exam components have been combined with a set of procedures to form the exam process used to assess promotion candidates for maintenance positions within the CDOH.

EXAMPLES OF THE PROMOTION PERFORMANCE APPRAISAL IN USE

1. SENIOR HIGHWAY FOREMAN

A total of 31 employees applied for promotion to this position representing all eight Maintenance Districts in the CDOH. The rescaling procedure employed with these candidates involved a determination of a grand mean from the 31 scores and adjusting the scores, by district, to the grand mean. This has the effect of placing all of the districts on the same rating scale with a midpoint equal to the grand mean. The district was employed as the unit of evaluation because SMEs and previous exams indicated that ratings within a district were comparable, but across districts there were often large discrepancies in rating scores.

Although leniency errors were suspected in the Senior Highway Foreman data, the small N prevented a systematic evaluation of the phenomenon. Therefore, the rescaling procedure was conducted on the initial scores received from the reviewers. Appendix C presents the data used in the rescaling procedure including the grand mean, district means and adjustments used for ratings within a district. Data are also included on the cutting score which was used to invite top applicants to an oral exam.

2. HIGHWAY FOREMAN

The same rescaling procedure was followed for the Highway Foreman with the exception of one step. The large N in the Highway Foreman exam (118) permitted a systematic evaluation of leniency errors. Appendix D presents the data used in the evaluation. Group 1 scores represent total rating points by reviewers which fell within the "acceptable" range of the Ratings Distribution Check Form. The grand mean for this group was equal to 32 with a range of 25 to 38. The scores in group 2 (27% of the total of 118) represent values outside the "acceptable" range. The grand mean was equal to 46 and the range was 42 to 49.

These data clearly indicated a leniency effect for the scores in group 2 and also demonstrates the range compression which often accompanies leniency errors. Following through with the plan for such rating problems, we resubmitted these scores to the reviewers with an identification of the problem and a request that the scores be altered to comply with our original guidelines. This request was accompanied by a letter of support from top management. It is important to point out that the rescaling procedures could have been applied to the scores as they were originally received (i.e., some reviewers following the Ratings Distribution Check form and some reviewers failing to follow this form); however, the future integrity of the process required a more direct approach. Following the resubmission of the inflated ratings, the rescaling procedure was implemented with the Highway Foreman ratings. Appendix E presents the data used in rescaling. The table also indicates the results for the ratings where the reviewers failed to bring their ratings within the "acceptable" range even after they were given a second opportunity. As can be seen in the column labeled "number to exam", District VI was represented in the exam in a manner comparable to other districts of its size (District I) despite the widespread uncorrected inflation of ratings. Based on the cutoff score, the top applicants were invited to a written essay exam.

Although only a limited amount of validity data is available for the PPA, the data that has been obtained is very positive. For the Senior Highway Foreman position, a criterion measure was developed from the oral exam data. The following results were obtained by correlating PPA scores with the oral exam results:

$r = .45$ $t11 = 1.67$, p less than .1 (one-tailed, uncorrected)
$r = .58$ $t11 = 2.38$, p less than .025 (one-tailed, corrected)

Similar results were obtained from the Highway Foreman position using the written essay exam results as a criterion measure.

$r = .30$ $t36 = 1.88$, p less than .05 (one-tailed, uncorrected)
$r = .39$ $t36 = 2.54$, p less than .01 (one-tailed, corrected)

Finally, a validation study was conducted for an Engineering Technician position using a nationally standardized test as a criterion measure.

$r = .33$ $t26 = 2.05$, p less than .025 (one-tailed, uncorrected)
$r = .43$ $t26 = 2.42$, p less than .025 (one-tailed, corrected)

These results are very encouraging as far as the validity of the PPA is concerned; however, future validation efforts will concentrate on the acquisition of objective on-the-job criteria data and data which more closely follows a predictive validation paradigm.

IX. SUMMARY

To summarize, recent exam procedures used to identify promotion candidates for maintenance type positions were found to be inadequate for reasons including efficiency, a lack of sound psychometric problems, and a lack of consistent application. An exam process was developed and implemented to alleviate a number of identified problems. The exam process was primarily based on a promotion performance appraisal that allowed for job performance to become a major factor in promotion decisions. In addition, for those positions requiring a second level of evaluation, an item bank was developed to provide ongoing information for building the evaluations. Data were presented on the validity of the PPA and examples of the exam process were presented to demonstrate implementation procedures. Overall, the exam process was found to alleviate many of the recently encountered problems with maintenance type positions in the CDOH.

References

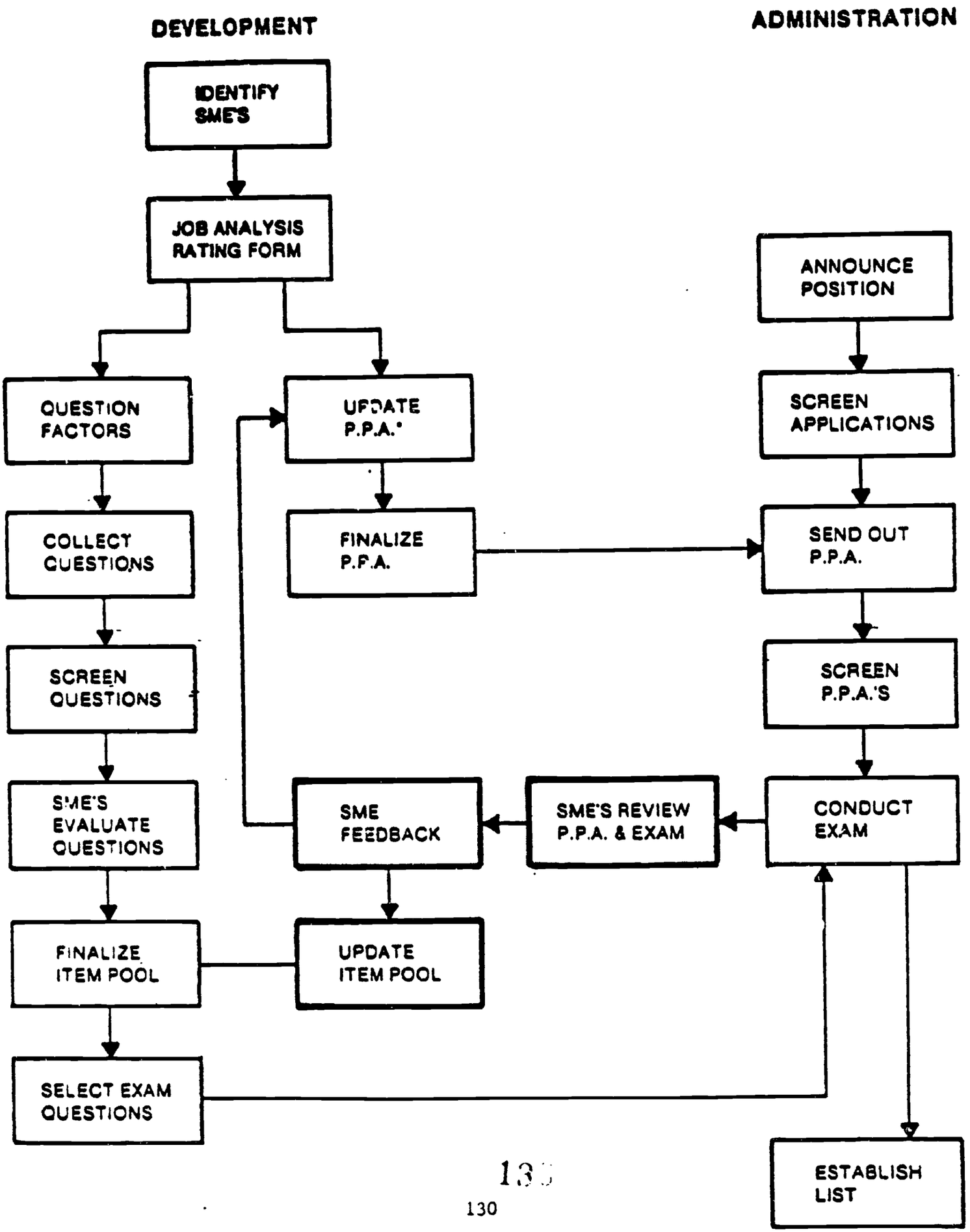
Mabe III, P.A., & West, S.G. (1982). Validity of self-evaluations of ability: A review and meta-analysis. Journal of Applied Psychology, 67(3), 280-296.

Maher, P.T. Departmental Ratings for Promotional Examinations. Paper presented at the meeting of the International Personnel Management Association Assessment Council. New Orleans, June, 1985.

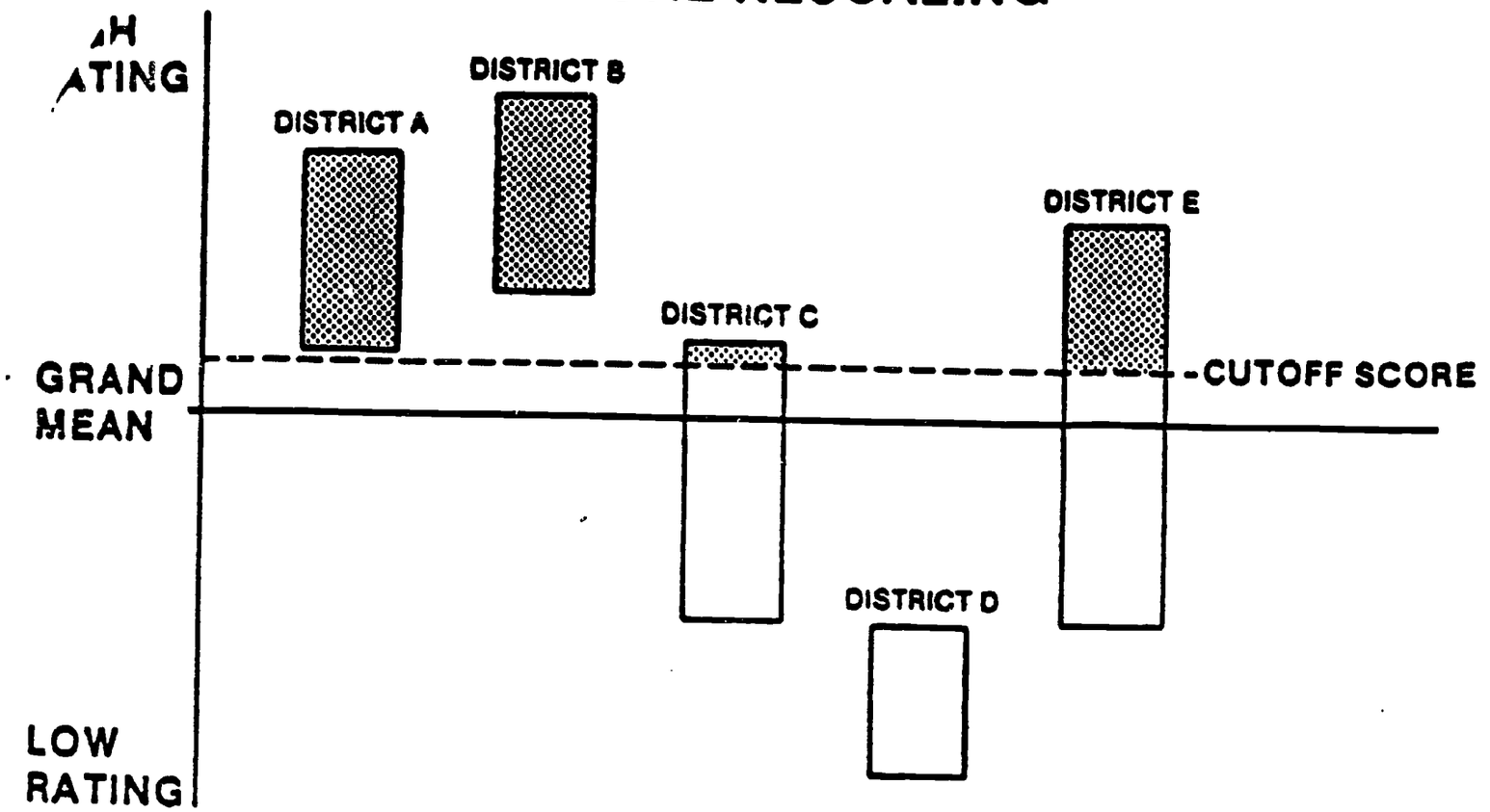
Nunnally, J. Psychometric Theory. New York, McGraw-Hill, 1978

APPENDIX A

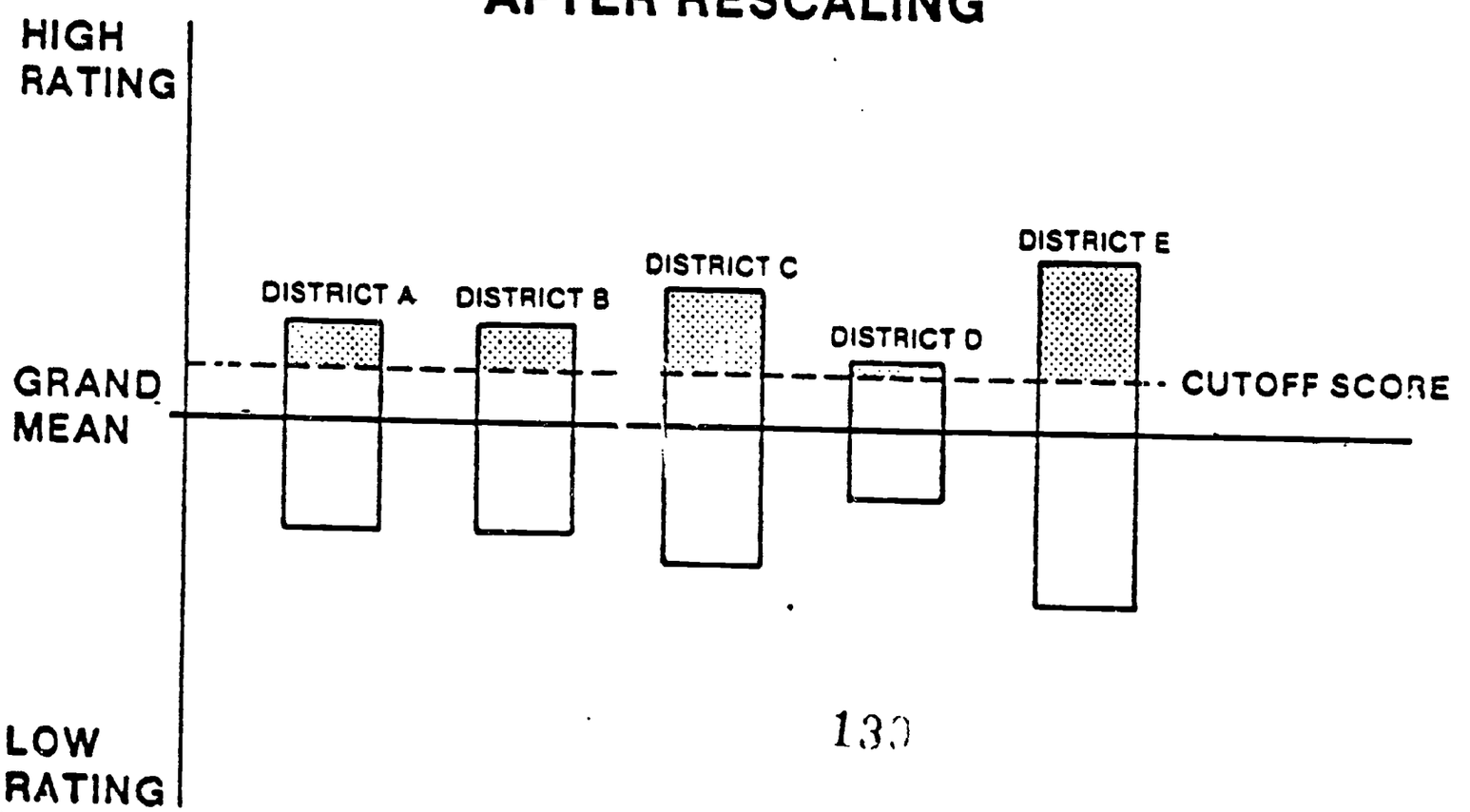
EXAM PROCESS



BEFORE RESCALING



AFTER RESCALING

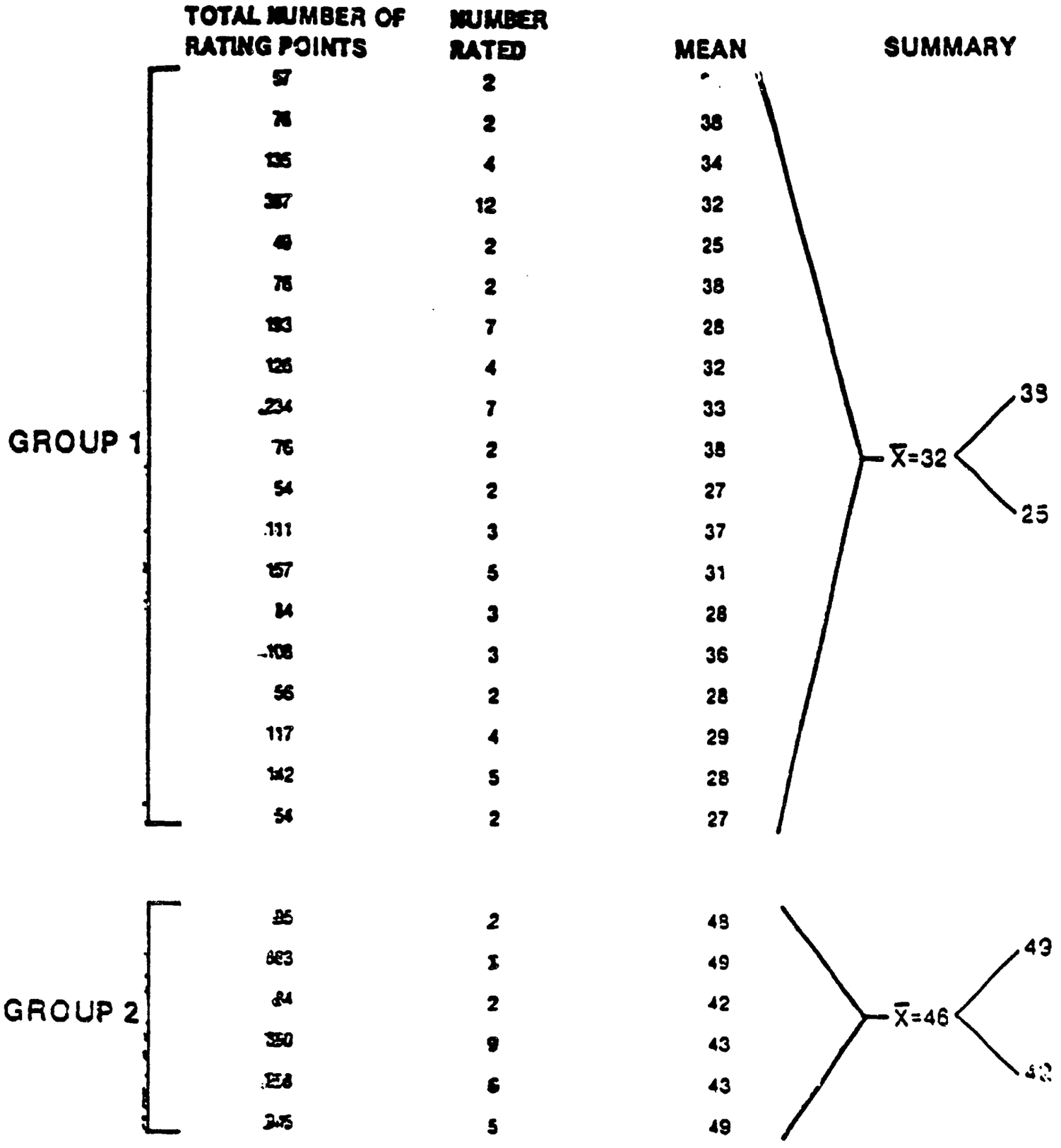


SENIOR HIGHWAY FOREMAN DATA

Grand Sum = 1,167
N = 31
Grand Mean = 37.65
Cutting Score = 38.00

DISTRICT	N	SUM	MEAN	ADJ.	# to EXAM	%
DIST. 1 (Aur.)	5	168	33.60	+4.05	2	40
DIST. 3 (G.J.)	4	134	33.50	+4.15	1	25
DIST. 3 (Crg.)	2	50	25.00	+12.65	1	50
DIST. 4 (Gre.)	7	258	36.86	+.79	4	57
DIST 5 (Dur.)	5	208	41.60	-3.95	2	40
DIST. 5 (Ala.)	2	65	32.50	+5.15	1	50
DIST. 6 (Den.)	2	84	42.00	-4.35	0	0

LENIENCY ERROR DATA



HIGHWAY FOREMAN DATA

Grand Sum = 4,108
N = 118
Grand Mean = 34.81
Cutting Score = 36.00

District	N	SUM	MEAN	ADJ.	# to EXAM	%
DIST. 1 (Aur.)	22	681	30.95	+3.86	7	32
DIST. 2 9 Pue.)	13	494	38.00	-3.19	4	31
DIST. 3 (G.J.)	13	415	31.92	+2.89	4	31
DIST. 3 (Crg.)	9	310	34.44	+.37	3	33
DIST. 4 (gre.)	17	612	36.00	-1.19	5	29
DIST. 5 (Dur.)	9	283	31.44	+3.37	4	44
DIST. 5 (Ala.)	13	431	33.15	+1.03	5	38
DIST. 6 (Den.)	22	882	40.00		6	36
		4108	34.81		40	

* Separate adjustments for reviewers falling outside the 'acceptable' range and failing to correct rating scores

N	SUM	MEAN	ADJ.
9	390	43.33	-8.52
6	258	43.00	-8 19
7	234	33.43	+1 38

Computer Assisted Proctoring: A Better Way to Administer Tests

Theodore S. Darany, San Bernardino County Personnel, San Bernardino, CA

This paper proposes the development of the capability to administer tests through a computer process. This process will be called Computer Assisted Proctoring or CAP. Some detail will be provided to explain what CAP is, why it is beneficial, and how it may be implemented in a practical setting.

Computer Assisted Proctoring: What is it?

Computer Assisted Proctoring (CAP) refers to the administration of tests by means of a computer. The computer includes such elements as the central processing unit, color display, audio feedback unit, printer, keyboard, program and data storage, and a telephone connection. This section addresses four elements of CAP: 1) Counseling and Intake, 2) Test Administration, 3) Test Scoring, and 4) Feedback.

Counseling and Intake: CAP can play a useful role in the initial contacts with those wishing to take an examination by serving as a counselor and intake specialist. The computer can provide the job seeker with a list of the types of jobs the civil service or personnel department currently has available for examination and corresponding job requirements. In turn the computer can obtain background information on the potential candidate. If the individual wishes to take a specific examination, the computer could administer it "on the spot." If not, the computer may retain the background information obtained in the form of an application to be forwarded to the personnel department for review and scheduling of the examination at a later date.

Test Administration: When CAP is administering a traditional test, it will offer it question-by-question to the candidate. This will enable the candidate to study the question, respond to it. In addition, he can indicate whether or not he would like to review that particular question later during the examination time period. At the end of the test period, for this traditional test, the candidate would again be presented those questions he previously earmarked for later review. At that point he may elect to change any answers. If time permits, he may also be allowed to review the entire test and his responses to each question.

If instead of a traditional test CAP is administering a "speeded" form, such as a name and number comparison test, it can more tightly control the administration of questions and the amount of viewing time for the section of questions as a whole.

Alternately, CAP can control the presentation of this speeded test item by item rather than controlling the duration of the administration of the entire test. The test designer can then more closely derive from the test exactly what it was designed to do.

The third alternative presentation mode which CAP permits would be that of a computerized adaptive test or "tailored" test. In tailored testing, the computer administers ability or aptitude test questions and successively computes approximations of an examinee's ability on the attribute being measured. In principle, the computer's estimation of the examinee's ability progressively becomes more and more refined after the administration of each test question. During this process the computer offers the examinee the single most appropriate question for that candidate that is available in its bank of test questions, which will tell us the most about the examinee's ability within the characteristic being assessed. In short order then, CAP can come to an estimation deemed satisfactorily accurate by the requirements preset by the test designer. At this point, our CAP simply stops administering questions, as no more are needed for accurate measurement. The tailored testing approach gives an accurate estimate of the individual's ability by using as little as 20% of the examination time as compared to the more traditional type of test.

Test Scoring: CAP can score the test during its actual administration, thereby providing an immediacy of score results for both the personnel department, and the candidate unavailable with traditional test administration.

Feedback: CAP can provide immediate feedback on the testing session to the candidate, including a simple test score display or significantly more. The range of possibilities includes comparison with other examiners, analysis of strengths and weaknesses, and suggestions for further training. CAP can also be programmed to initiate any of several forms of "on the spot" training.

Computer Assisted Proctoring presents a number of positive attributes: 1) Cost Effective, 2) Tireless, 3) Consistent, 4) Efficient, and 5) Versatile.

Tireless; Consistent; Efficient: Unlike human test proctors, CAP does not tire and is always ready to administer a test. CAP has the ability to administer a test consistently every time and thus brings new meaning to the term "standardization" in testing.

CAPs efficiency is derived in a number of ways. The first is the previously mentioned capability to offer tailored tests. Reducing length of tests not only makes testing sessions more productive, but reduces measurement error due to examinee fatigue as well. Other benefits of CAP include reduced opportunity for cheating in large group settings, the possibility of offering alternate forms of the same test content, and the ability to offer the same test items in a variety of sequences.

CAPs efficiency is further extended by the relative ease of maintaining permanent CAP facilities in locations remote from a central civil service or personnel agency. Such facilities would enable the civil service or personnel agency to offer dramatically improved service to persons in outlying regions at a lower cost.

Versatile: When properly equipped, including a color display and high resolution video graphics, an audio system with voice synthesis capability, a printer, sufficient storage capacity, and a telephone communications device, CAP offers tremendous versatility both to the central personnel agency and to potential examinees. The disadvantages faced by the handicapped in a traditional test setting can be overcome through the use of CAP. The computer can administer instructions and test questions visually for deaf examinees. The blind can receive the test in audio format with a verbal confirmation of all information/answers typed in by the examinee. With a computer set up for speech recognition, the paraplegic examinee who cannot operate a keyboard will still be able to proceed through the examination receiving either audio or visual response from the computer.

The final and most intriguing benefits which might result from Computer Assisted Proctoring derives from the computer's ability to simulate actual situations related to the job being tested for. The ability to graphically display outcomes to responses given by the examinee would provide a much more accurate and efficient tool for assessing complex problem solving, analytical and decision-making skills than is currently provided by paper and pencil tests or multi-day assessment centers. Such simulations could also be administered in a series to test the candidate's ability to learn from mistakes and benefit from the correct decisions made along the way.

Cost Effective: There are a number of potential dollar savings available through the use of CAP. Among them are:

1. reduced cost of paper and forms, as well as forms handling, storage, and mailing.

2. reduced payroll cost for proctoring.

3. reduced scoring costs due to the elimination of the answer sheet scanning process.

4. reduced staffing costs due to increased efficiency in dealing with handicapped examinees, reduced need to maintain a large staff to deal with flexible test administration requirements, and reduced chance of appeal or grievance due to scoring disputes.

How May Computer Assisted Proctoring Be Implemented? CAP may be implemented through terminal access to a large main-frame computer or through any of several currently available micro-computers. Major requirements for either system would include high quality graphic and character displays, attached printer, speech synthesis or high speed random access to audio tape segments, speech recognition, and sufficient speed to allow proctoring of the examination without undue delays in the computer's own response cycles.

Conclusion

Computer assisted proctoring offers a number of sizeable advantages to test administration. The advantages range from obvious cost beneficial consider-

ations to providing better service, and to providing services now practically impossible. CAP can be implemented on computers ranging from large main frames to micro-computers. The choice of the type of computer should focus on the implementing agency's current capabilities relative to hardware and programming skill as well as several demographic factors such as distances between examination centers. Either approach could be practical in a given setting.

* * *

Computerized Simulation Testing: A BASIC Language Program to Develop and Automate Simulation Tests

Larry S. Jacobson, Connecticut State Department of Personnel, Hartford, CT

INTRODUCTION

The following paper briefly describes a BASIC language program that is being developed to assist in the construction and administration of "computer simulations." The program to be described allows users to enter written simulations (latent image) or other simulation oriented tests into a microcomputer.

This paper also includes a discussion of advantages and disadvantages of computerizing such exams as well as some future directions computer assisted testing might take.

Background

Simulations, depending on one's definition, have been around for quite some time. One of the earliest examples of a "work sample simulation" exam was conducted by German and British Psychologists. These researchers recognized that for critical, complex or high level positions, written examinations had little predictive value in selecting candidates. For example, British psychologists set up a 3-day house party, where civil service candidates were observed by trained assessors. Their results indicated that "assessor" judgments were superior to that of written exam in predicting later job performance.

Simulations have continued to develop along a number of paths. For example, there are "role playing" simulations which are used within the context of an oral exam, where a candidate is provided with background information and asked to assume a role. An actor or the oral panel then confronts the candidates with realistic problems.

Simulations are also found in the context of assessment centers where assessors observe and rate a candidate's performance in simulation exercises. An elaborate demonstration of such a procedure was broadcast on CBS's "60 Minutes" several months ago. This assessment center procedure involved the staging of a terrorist-hostage negotiation situation complete with gun fire, medical emergencies, and high stress problem solving situations.

Unfortunately, many of us in the public sector seldom have the opportunity or resources to conduct more elaborate simulations. However, for some time (at least the early 50s) the written simulation has been used extensively in training and licensure primarily in the health care profession. By contrast, very few instances of non health related simulations have been cited in the literature, with a few coming from areas of teacher education, rehabilitative counseling and public safety.

Written Simulations

Briefly, a written simulation test or exercise usually involves giving a candidate a hypothetical problem with some background information. To solve this problem the candidate must make a number of choices which involve gathering information, following directions and selecting courses of action. What distinguishes this approach from other exam modes is that the actions candidates take result in feedback about the consequence of their response. Further, unlike a multiple choice type exam, once a response has been made, it cannot be retracted.

The primary method used to administer written simulations is with a latent image procedure (not to be confused with the latent trait procedure). In this approach candidates are posed with a problem and may select from a number of possible options. Once a option is selected the candidate rubs a specially treated marker across a specified area of the answer booklet. The chemical from the marker causes a preprinted "invisible" ink to become visible and reveal further information to the candidate. This information can include directions informing the candidate to proceed to another section, further information about the solution to a problem, or feedback about the consequence of some action.

The purpose of the present paper is to describe a BASIC language program package being designed to assist in the development and administration of computer simulation problems. Computer administered simulations have a number of advantages over the written method. And, even if confined to the written mode, the following program will be useful in the development of written simulations.

Use of this program, however, does not reduce or eliminate all concerns associated with simulations. For all practical purposes computerized simulations will require some painstaking approaches to job analysis, the same necessity for using motivated and imaginative Subject Matter Experts, and similar difficulties determining psychometric quality.

SIM-U-PLAN

The program to be described was created by Bruce Davey and is still being developed and tested. However, the main "engine" of the program has been completed and was utilized to convert two written simulations to the microcomputer. SIM-U-PLAN has been developed along the same lines as other generic program packages such as spread sheets or word processing packages. First, it is flexible enough to handle many of the branching features characteristic of simulations. Secondly, use of this program takes very little knowledge of BASIC language and requires the user to learn a vocabulary of fewer than 10 words.

Program Functions

SIM-U-PLAN is composed of three major programs:

SIMLOAD - A simulation largely consists of narrative text. SIMLOAD takes this text and loads it into disk files (hard disk or floppies) which are later used by program SIM.

SIMUTEST- Once text has been entered by the SIMLOAD program SIMUTEST tells the user if the program has loaded properly and alerts the user if the text does not properly fit on the computer screen.

SIM - SIM is the "generic" simulation runner, designed to run most simulations. It makes use of meta-language (a simple natural vocabulary) which provides directions to the program for starting, stopping and branching the simulation.

SOME FUTURE PROGRAM ADDITIONS

SIMWRITE- A wordprocessor type program will be developed that will allow a user to more easily enter text into the simulation database.

SIMDEBUG- A diagnostic program will be written which will read through a simulation to detect logical/structural errors (e.g., determining that user has branched to a nonexistent section, or the rules of the meta-language have been violated in some way.

Program Operation

The user writes out the simulation problem as a series of DATA statements. Inserted within these DATA statements are SIM (meta-language) words or directions. These words are considered a meta-language rather than a programming language because it runs out of BASIC; in other words, BASIC serves as its interpreter. Consequently, the user need only be concerned with the SIM(ple) vocabulary and not BASIC language itself. (An illustration of a short simulation was presented).

Using SIM-U-PLAN

To date, two simulations have been adapted using the present simulation package. First, a sample vocabulary patient management program (courteously provided by the Professional Examination Service) has been automated utilizing SIM-U-PLAN. Secondly, a "mystery" type problem entitled, "The Teacher and the Threat" (supplied by Bruce Davey) has been adapted.

Both simulations, although very different in original format, required little modification for computerized simulation. The majority of effort went into the inputting and formatting of text for proper computer display.

Our initial experiences with computerized testing suggest that a microcomputer would be preferable to the risk of system malfunction and lack of immediate test administration control offered by a terminal tied to a mainframe or minicomputer. We have found that the micro will present simulation materials to candidates at sufficient rates of speed. Further, PCs with only floppy disk storage usually have sufficient capacity to store moderate sized simulations (although a 10 or 20 meg Hard Disk is recommended if the simulation uses a large number of problems).

As with written simulations, special care should be taken in providing candidates with sample problems prior to actual test administration. Few candidates have taken computer administered exams. Consequently, test developers will have to deal with a number of the following concerns:

1. Is the candidate comfortable having the exam administered by a computer?

2. Have candidate responses been simplified as much as possible? In other words, candidates should be required to only make simple keyboard responses.

3. Is the exam administration "user comfortable-amiable?" Does the candidate have the opportunity to refer beck to earlier materials? If not, has appropriate information been printed out? Is the material formatted on the screen in an easy to read manner? Are the exam displays paced, or is text frenetically "flashed" on the screen?

4. Has the program been "bullet-proofed" to prevent as many input or system crashes as possible?

We will require more research and actual testing experience before we have a clearer picture of the impact of computerized versus written exams on candidate performance.

Advantages and Disadvantages of Computerized Simulation

Although we have not yet had the opportunity to try the computerized approach on "real" candidates, several positive and negative features of such an approach have become evident.

Advantages

Training a candidate to take a computerized simulation exam should be easier than with the latent image procedure. More of the branching directions are handled by the computer, simplifying directions to the candidates.

Special "invisible ink" printing is not required. In fact, depending on the nature of the problem, the computerized simulation could eliminate most of the paperwork and printing associated with a written exam.

The computerized simulation allows more assessment flexibility, for example, certain sections of an exam could assess the candidates use of time, and resources, as well as strategies used in solving problems. (A written simulation can also accomplish this but requires greater administrative and scoring effort).

As discussed earlier in Bruce Davey's paper on non-cognitive testing, the computerized simulation could be designed to detect candidate response inconsistencies. If such response inconsistencies could be revealed to the candidate without "giving away" simulation answers, exam reliability could be increased.

The SIM-U-PLAN package should make it much easier to develop, edit and modify simulations regardless of the final administrative mode.

Pilot administrations of the computerized simulation approach suggest faster administration time than with written simulations.

As mentioned earlier, written simulations do not allow for the retracting of responses. The computerized simulation would help reduce accidental responses by giving the candidate one more chance to "SELECT AGAIN," before a final choice has been made. This could reduce some response errors.

Disadvantages

Having your microcomputer or terminal "go down" is not the same as having a defective test booklet. Backup hardware and software will be required to ensure against major computer malfunctions. Unfortunately, unlike other written exams, the candidate cannot retake the same examinations should a problem develop. However, if a backup storage device such as tape or disk were utilized, it might be possible to restore candidate responses up to the point at which the computer malfunctioned.

The Future

There are several technological developments that paint an optimistic picture about the future of computerized simulation testing. For example,

recent developments in compact disk/optical disk (CD) will provide sufficient memory size and speed to allow the presentation of video information at an acceptable rate of speed. In contrast to a candidate reading the simulation, and making simple keyboard responses, the candidate will be able to make verbal responses and be able to see and hear the consequences of their actions. The technology for both voice input as well as video retrieval are already available.

* * *

A Computer Administered Interest Inventory

Bruce W. Davey, Connecticut State Personnel Division, Hartford, CT

The microcomputer has tremendously changed not only what goes on in personnel assessment but many aspects of our way of life in the past five or ten years. Witness the fact that TIME magazine named a computer as its man of the year a few years ago. Some people were upset about that, but I could not think of a more appropriate choice.

In particular, computers have changed the way we gather and analyze information, and the way we crunch numbers. And what else is testing, but the gathering and analyzing of information about people and the conversion of that information into numbers? In the recent past, microcomputers have tremendously enhanced our ability to perform those tasks.

But as Larry Jacobsen said earlier, a society's tendency to make full use of new technology tends to lag behind the availability of that technology. And that's the way it's been with testing and the microcomputer. We haven't begun to take full advantage of its capabilities yet.

Microcomputers have the capability to make the tests we give much more sophisticated and interactive and efficient than they presently are. And yet, for the most part, the tests I've seen transferred to the microcomputer haven't come close to taking advantage of the machine's capabilities. Far too many of them look like nothing more than paper and pencil tests flashed on a computer screen. I consider that to be a waste of the computer's potential and power.

For about the next fifteen minutes or so, I'll be talking about personality and interest inventories. I hope to demonstrate how they can be "jazzed up" a little bit to better take advantage of this new microcomputer technology.

I'd like to start this discussion by outlining some of the capabilities of the microcomputer which I think can enhance the testing process, whether we're considering personality and interest testing or other types of tests.

One major capability of the computer is its ability to interact with the candidate. In a way, it can talk to the candidate, and it can tailor its conversation to what that candidate is doing. It can call the candidate by name. It can stop the show and tell him when he's made an impossible response or done something that needs correction. It can provide feedback--something that candidates very much want but that we testers have never thought much about, out of necessity. Now we can think about it. And the computer can even monitor the candidate's response consistency and point it out to him if he responds markedly inconsistently. That's something I'll talk about later.

A computer can allow for much freer response possibilities, it's not artificially confined to five choices by the size limitations of a machine--scorable answer sheet; so if you want to offer candidates ten choices to choose from or if you want them to make ratings on a fifteen point rating scale, the computer can accommodate.

A microcomputer can also ask the candidate to make more than one choice in responding to an item. If you have much experience writing test questions, you probably recall many time where you'd like to have keyed more than one choice as correct, or where you'd like to have included a number of choices and asked candidates to select as many as they think are correct. This is easily done with a microcomputer.

Also, within limits, it can request and grade free responses. For example, it can ask you who our third President was and be prepared to give credit for Jefferson or Thomas Jefferson or maybe anything that ends with Jefferson or even a reasonable facsimile. Or it can present a math problem and request a correct answer, and mark it right if it's in whatever is set as the acceptable range of tolerance.

A microcomputer capturing a candidate's responses can also perform sophisticated mathematical or analytical operations practically instantaneously as the candidate makes his or her responses. That allows for tailored decisions to be made. such as Ted described when he talked about computerized adaptive testing. When I describe the computer-administered inventory I'm going to be talking about, I'll talk about some other sorts of tailored decisions a computer can make in monitoring AND TAKING STEPS TO CORRECT a candidate's response inconsistencies.

In a vein similar to computerized adaptive testing, the computer could calculate how internally consistent a candidate's responses have been on a particular subtest or on some homogeneous scale, and if the reliability is shaky it could give that candidate additional items until the reliability level was acceptable. This becomes especially feasible if the items of the test are calibrated something like they are in computerized adaptive testing, so that you can fix a performance level with fewer items. That's admittedly a lot harder to do with personality and interest tests because

you're not working with pure abilities--but I think the methodology applies and is feasible. For example, take the characteristic of aggressiveness by giving them an item which is calibrated somewhere in the middle of your aggressiveness scale--maybe something like, "If your steak isn't cooked the way you wanted it, would you send it back?" If they answer yes, they get an item calibrated at a higher level of aggressiveness; if they answer no, they get an item at a lower level. And I'd bet you'd have a fairly stable fix on this person's aggressiveness score within about a dozen or so items.

Now, I'll start to talk about the computer-administered interest inventory.

The test in question is called the Vocational Interest Questionnaire--or VIQ for short. It was developed by an eccentric named Bruce Davey--and the philosophy behind the test is worth discussing here. The VIQ was developed as a reaction to interest inventories like the Strong Campbell Interest Inventory, which asks you hundreds of questions related to narrow activities such as writing letters or watching parades or baking or things like that. From that they compare your interests with those of people in particular occupations, and score you on how similar or dissimilar your interests are to those of people in those occupations.

The VIQ is a reaction to that because it's only 52 items long rather than 400. But they're 52 items that are each designed to be broad and meaningful in their own right, so that I think the 52 items cover about as much ground as the 400 items. It's easier for candidates to complete the test; I think the responses are much more stable because the items are much more meaningful; and the results can be interpreted almost clinically. In addition, the profile of the person completing the VIQ can then be compared with job profiles for other jobs to not only show a person what kinds of jobs their interest pattern best corresponds to, but to show them why. For example, if you have a person before you who thinks he wants to be a computer programmer, you can compare the two patterns fairly directly and easily and tell them why they don't match up. You might say something to him like, "Well yes, you like machines, but you don't like math and you have only average interest in intellectual challenge and in detail work--and those are very important for programmers."

In its non-computer-administered form, the VIQ has 52 items which the test taker rates on a five point Likert scale ranging from "Like Very Much" to "Dislike Very Much." I had always been troubled by what I would expect to be a fairly low Test-Retest reliability for this test, in part because it's short and in part because the rating scale isn't that well articulated. However, we forged onward. In Connecticut, we administered the VIQ to about 800 State employees in three separate validation studies and to maybe 5,000 candidates. It was successful in all three validation studies, by the way. So--there was a useful database available on this paper and pencil version of the VIQ. That usually encourages test developer to stick with the original version that has all the normative information.

But I could see ways to improve the VIQ in a microcomputer-administered environment, and to maybe solve my problems of dubious retest reliability--which incidentally I still had no data on.

So let me show you what the VIQ looks like now. First of all, when you sit down with the machine, it captures basic information about you--your name, sex, race, occupation, and educational level--and then it gives you a very brief description of the test. With little further ado, it moves into test administration.

Each VIQ item is presented one at a time. Appearing along with the item on the screen is a fifteen point rating scale. That's one way in which the VIQ takes advantage of its computer-administered nature--it uses an extended rating scale for finer discrimination.

Figure 1 shows how the VIQ item and rating scale appear to the candidate on the computer screen.

Another feature of the computerized version of the VIQ is its ability to measure and monitor the candidate's consistency, and to actually improve it. It does this by re-administering the entire test to the candidate. Figure 2 shows how the computer introduces the second administration of the test. Remember that this is only a 52-item test that only takes about 15 minutes to complete.

It may seem to you to be a bit of a nuisance to run the poor candidate through the same test twice, but some major benefits accrue from doing so. You can now monitor the consistency of individual candidates, and if they <u>are</u> inconsistent, you can tell them so, and force them to give more thought and care to what they're doing.

Let's assume that Ted Darany is our test-taker, and on the first VIQ administration he rated the "Artistic" item as a 9. On the second time through, he has a stunning difference in ratings, but it's wide enough to express concern over, and that's what the computer does--it compares the two responses and stops the show. Then it points out Ted's inconsistency to him and asks him to think deeply about this item and try again. You can see the text for yourself.

You will note that at the point an inconsistency is detected, the computer tells Ted that his <u>next</u> response is the only one that counts. The idea here is that with two ratings which differ from one another, one of them might be just plain wrong and therefore it shouldn't be averaged in. My present feeling is that be telling the test-taker he's being inconsistent and asking to carefully reconsider, that last response is the best and most accurate one. Accumulated research may eventually prove me wrong. Maybe the research will show that I should be taking the average of all three ratings--at which point I'll revise the way such items are handled in scoring...but for now, I like this approach.

Another way in which the candidate's consistency is monitored is that the candidate's two sets of ratings are correlated with one another and included in the final report. By recording and reporting this critical piece of data on the candidate's consistency in responding, we are giving the evaluator key information on how trustworthy this particular candidate's

responses are. The evaluator can consider the results not only from the standpoint of the reliability of the test instrument, but also from the standpoint of the reliability of the individual test-taker's responses.

For all VIQ items other than the "time-out" items, the final rating of each item is the average of the two administrations. That also means you can calculate the reliability of the final ratings by taking that correlation between the ratings and plugging it into the Spearman-Brown prophesy formula. The computer does that. And so far, the reliability of the individual candidate profiles is running about .90 or .91 on the average. Unfortunately, that's based on only 22 people who have taken this fairly new version of the VIQ. The range of those correlations, by the way, goes from .688 to .966, and the range of reliabilities therefore goes from .74 to .98--which isn't bad for a 52-item interest inventory.

TABLE 1

A. INDIVIDUAL TEST-RETEST RELIABILITIES FOR THE VOCATIONAL INTEREST QUESTIONNAIRE

(NOTE: Based on only 22 people)

	Correlation Between Test and Retest	Reliability: Test & Retest Combined
Lowest	.588	.74
Highest	.966	.98
Median	.815	.90
Mean (using Fisher's z)	.840	.91

B. INCREASE IN RELIABILITY DUE TO REPEATED ADMINISTRATION

Conventional Version Reliability	Reliability if Administration is Repeated
.60	.75
.70	.82
.80	.89
.90	.94

And I should point out that these aren't scale reliabilities but are based on r between rank-orderings of all items...

I'd like to make some comments about that reliability figure for you to mull over so you can decide whether it's an overestimate or an underestimate.

One thing that argues for it being an overestimate is that some of the correlation between the halves may be due to somebody remembering their first responses and some of it may be due to mood factors that carry over through the two administrations and won't be there tomorrow. In other words, if the two administrations were separated by a larger time interval, that correlation is likely to be lower.

One thing that argues for the reliability figures being an underestimate is that it's between the two halves of the test without figuring in the positive effects of the computer's intervention. In other words, I think that the computer's intervention to point out inconsistencies and give the candidate a chance to reconcile them greatly increase a candidate's response consistency in a way that isn't reflected in the correlation between the halves...and for some candidates that makes a major difference (the one at .588).

This consistency monitoring will make the biggest impact where it's most needed--people who were inconsistent will be boosted the most.

This machine, the microcomputer, provides some opportunity for new levels of creativity and better measurement and I hope were going to use it to its fullest over time.

Ted has such a good line during his talk and I'm going to import it to mine. We're here today as cheerleaders to get you to consider and find ways to apply this new technology. Today we're presenting possibilities and developing applications. In the immediate future I hope we'll be seeing not possibilities but established applications.

* * *

ORAL EXAMINATIONS: UNIQUE APPROACHES TO DEVELOPMENT, RATING SCALES AND RATER TRAINING (Paper Session)

Development of a High-Structured, Competency Based Oral Exam for Police Sergeants

Bruce Davey and Karen Duffy Wallace
Connecticut State Personnel Division, Hartford, CT

It is a well documented fact that the unstructured interview has a very low level of validity. Recent studies, however, have shown that structured oral examinations have good levels of validity. This raises a question: since adding structure to the interview increases its validity, does a very high degree of structure lead to still higher validity?

In a paper presented at the 1984 IPMAAC Conference, one author concluded that even with a fairly high degree of structure in the oral examination process, oral panels tend to systematically differ from one another in terms of average score levels, variance, cues attended to, and validity of the final results. The present authors attempted to address this problem of differences across oral panels by building in a very high degree of structure. This was done in order to assure the fairness of promotional oral test for State Police Sergeant, administered to 352 candidates by three separate oral panels.

Since this was the first time recently that oral exams had been used for Sergeant there was a lot of grumbling. The union threatened to enjoin the exam process, (didn't carry it out), and Troopers were reported to compromise the exam by posting a list of the questions after the first three days.

Several security measures were taken to safeguard the exam process. Candidates had to sign statements indicating no knowledge of the exam, and under oath to indicate that they would not discuss the content until all candidates were examined. Examiners were also sworn to secrecy about the exam questions.

A thorough job analysis was conducted covering tasks performed and the KSA's needed to be successful at the entry-level. Job analysis questionnaires were computer scored and analyzed to focus on test development efforts. Considerable attention was paid to question development with input from officer volunteers. Ratings were made of the importance of each question and 10 were selected from a larger number. Answer keys were carefully developed and score weights were assigned to detect highly specifically stated errors--weights of a minus nature on basis of importance and criticality of the error.

Three highly specific scoring keys were developed for each of 8 situational questions, each having a number of questions related to it, for 32 questions in all. Each key consisted of a list of elements candidates were expected to include in their responses, with specific point deductions if omitted. The point deduction scheme was tied to a scaling procedure which made it possible to tie candidate responses to performance levels and to establish a competency-based pass point.

Two other questions were scored using the more common approach of rater judgment using a Likert-type scale rather than a scoring key. This procedure was followed for eight of the ten questions. The final two questions concerned the candidate's interest in being a sergeant and preparation for that rank. Guidelines for the last two questions were not highly structured. These latter two questions were found to be highly susceptible to halo effect and to differences in means and variances across panels. The more highly structured questions were far less prone to these problems.

Two weeks prior to the candidate's exam date, he/she was able to pick up a 300 page study guide. Four hour training was conducted for all examiners, alternate and monitors. (Monitors were assigned to each of the three

boards). Candidates could exclude up to two examiners from the compilation of their final score. About 15% of the candidates chose to exclude 1 or 2 examiners. The score sheet allowed examiners to document a candidate's answers by making a check mark next to answers hit or missed.

Results

The reliability figures for each of the three panels were extremely high. With the amount of structure we introduced into the examination process, you would expect very close agreement. Still, we were pleased to get an average correlation between raters of .95 and a reliability of .99 for the typical panel. I should add, here, however, that these reliability figures aren't pure--they are based on the raters' final ratings, and raters were permitted to change their ratings after discussion. However, we estimate that raters changed their scores after discussion only about 10% of the time, and then the change was usually a change of a single point, or possibly two (See Table 1 at the end of this report).

We were also pleased that the mean scores and standard deviations for each panel were so close together. However--the means and standard deviations were not so close as to be interchangeable. There were two significant differences. Firstly, panel #2 rated about two points higher on the average than panels 1 and 3, and that was a significant difference at the .01 level. And secondly, panel #1 spread its scores out more--they had a significantly higher standard deviation than the other two panels (.01 level). For that reason, we decided to standardize scores for each panel. The message here, we feel, is that if you use multiple oral panels, even a high degree of structure is not going to wipe out differences across panels.

Although reliability doesn't assure validity, we feel that in this case, such a high reliability shows that the raters were attending closely to our scoring key and to the answers the candidates gave--not to extraneous clues such as appearance, verbal skills, and so forth. This confirmed our firsthand observations of each panel's performance.

The committee members' scoresheets led to a rich documentation file. For each question asked of each candidate, we had five checklists indicating what points the candidate had and had not handled well, and the number of points deducted by each committee member for each error. Thus, any challenges could be met with a thorough file of documentation as to why each rater gave the grade that he or she gave.

You may be wondering how examiners feel about high structure, and whether they resent or resist it. We frankly expected that some of the might well be resentful or resistant. Actually, we were pleasantly surprised. After the examiners got the initial "hang" of it, they were very comfortable with the amount of structure provided and with the amount of documentation the process generated. All fifteen oral examiners stuck closely to the rules, as evidenced by the reliability figures.

We have since tried the high structure approach for a higher level job, that of State Police Captain. Here the questions were primarily strategic and administrative in nature, but again we had no trouble designing an effective scoring system, and again the procedure was well-accepted by the raters.

A direct test of the characteristics of high versus moderate structure were built right into this exam process. Whereas the first eight questions followed a highly structured pattern, the last two ratings were judgmental in nature. The last two questions dealt with the areas of work experience and interests and, at the time, we felt it best to allow these two questions to be scored by committee judgment.

What we found was that the average intercorrelation between questions 1 through 8 was only .24, which seems to indicate very low halo effect, if any at all. However, the last two questions, which had no point deduction scheme, correlated .75 with one another.

Why is there such a great susceptibility to halo effect in an oral exam? We believe that it is because candidates do more than just answer questions in an oral exam—they also transmit a wide variety of signals. Although we hope that our examiners are primarily influenced by the candidate's specific answers to job-related questions, we have to recognize that examiners are strongly influenced by many other signals—speed of response, voice tone, steadiness of voice, nervousness, degree of eye contact, posture, dress, and physical appearance—in short, all those things we hope will be ignored, but which never are. Examiners also seem to be influenced, in grading a present question, by how well or how poorly the candidate has answered previous questions.

We feel that high structure greatly curtails halo effect, and focuses the oral committee back upon the content of the candidate's responses, rather than the candidate's style. The examiners are focused to indicate on their scoresheet what kinds of concrete errors of omission and commission the candidate has made in answering the question. This leads directly to a final score.

One other aspect of this process which we feel contributed greatly to the reduction of halo effect is the practice of scoring candidates question by question instead of factor by factor. We feel that requiring raters to evaluate broad factors invites halo effect because it invites ratings based on overall impressions. On the other hand, requiring raters to evaluate the candidates specific responses to each question really minimizes the opportunity to inflate or deflate a rating based on global impression.

One effective point of closure would be a direct comparison of the construct- and criterion-related validity of this oral exam with its predecessors. Unfortunately, we don't have that for you today. However, in a few weeks we will have a chance to make a direct comparison—about 1,000 candidates examined by 8 committees using high structure, versus 900 candidates examined by 8 committees in 1985 using moderate structure. Perhaps that will be the subject of a paper at the 1987 IPMAAC Conference.

Other positive effects:

* Candidate feedback--more specific info than we have ever provide before on why candidates didn't do well.

* Capability to generate item analysis data. Since these exams are scored question-by-question, it is possible to generate a printout which shows the difficulty of each question; the extent to which it spread out candidate responses; and its correlation to the other questions asked. After such rigorous analysis, any question defects should be spotted and corrected.

* A defensible, competency-based passing point as required by the Federal Uniform Guidelines and by the Joint Committee on Technical Standards.

Note: Test had no adverse impact.

TABLE 1
BOARD MEANS AND STANDARD DEVIATIONS

	Mean	S.D.	N
Board 1	84.46	6.83	120
Board 2	84.77	5.31	118
Board 3	84.72	5.73	114

BOARD RELIABILITY DATA

	Average r between raters	Alpha Reliability
Board 1	.981	.996
Board 2	.930	.985
Board 3	.935	.986

* * *

Raising the Validity of the Oral Examination: The BOSS Technique

Roger Davis, King County, Washington

In a much discussed and debated article this year Hunter and Hunter provided the results of their meta-analyses on a number of test contents and formats used in employment settings, covering among other topics ability testing, job knowledge tests, assessment centers, and interviews. While many of the meta-analysts' conclusions are enlightening, and some controversial, one result they produced is something specialists have believed for a long time, that interviews typically have very low validity. According to the Hunters the true validity of interviews is little more than chance (r=.14).

Without arguing some of the tenets of the School of Meta-Analysis, such as the futility of small-sample criterion validity studies, or that the variance from their true correlation coefficients which you find in your local study is due to your error, the author of this paper finds and develops some recent research indicating that it is possible to raise the low validity of the interview procedure in certain situations through the use of the B.O.S.S. technique.

Let me review quickly some of the data on interview validity. In 1976, Dennis Huett reviewed many primary interview validity studies, going back as far as 1916. Huett failed to conduct his literature review as a meta-analysis study, but he did itemize 51 separate validation studies in which over 53,649 people were covered by predictor/criterion measures. The predictor is always the same, interviewing. Where comparable validity coefficients were reported, the simple average validity was about .2.

Let me recount quickly three examples. In 1947, John Flanagan reported a study of two groups of air force cadets. The combined sample size was 632. The job was pilot; the criteria were job performance ratings. For one group the validity of the selection interview was .06 and .13 for the other group.

In 1960, Campbell, Prien and Brailey reported an interview validation study of 95 clerical trainees who were interviewed by trained professional psychologists. The criteria were supervisory job performance ratings. Result: r=-.17. That's negative .17.

In 1969, Douglas Bray, one of the founders of the assessment center movement, reported two ATT studies involving interviews by psychologists. The criteria were assessment center ratings and 10-year salary progress. Bray reports only that about two-thirds of the obtained correlations were significant. For his sample sizes of 200 and 148 hires respectively, statistical significance is reached between the .13 and .17 levels.

The results the Hunters find for the quality of the interview are consistent with the results found by Huett. The difference between the .14 they reported and the higher .2 estimate could easily be accounted for through file drawer analysis and my deliberately rough estimation. I accept .14 as

most likely the mean true validity of the interview proper, and of any particular interview in the absence of evidence to the contrary.

Is there any evidence to the contrary?

Yes. There is positive evidence that the <u>situational</u> interview, is much different. Here are ten studies:

STUDY	AUTHOR	JOB TITLE	CRITERIA	r	n
1	Davis	Supervisor+	Job Performance	.41	30
2	Davis	Supervisor+	Job Performance	.37	11
3	Davis	Pol. Off.	Training Acad.	.18	64
4	Latham	Foremen+	Job Performance	.41	62
5	Latham	Clerical	Job Performance	.47	29
6	Latham	Linemen	Job Performance	.14	157
7	Latham	Hrly. Wkrs.	Job Performance	.46	49
8	Latham	Laborers	Job Performance	.33	36
9	Latham	Laborers	Job Performance	.39	20
10	Davis	Supervisor+	Job Performance	.25	22

weighted r=.28
N=480
s.d.=.11
weighted managerial r=.38
managerial n=125

This table does not pretend to be a meta-analysis; it is just a list of small-scale studies. The correlations are not corrected for anything, as the Meta-analysts' are. So when Hunter and Hunter reported a .14 for the interview, that's about the limit of what can be said for that procedure.

Let's look at the situational data, and I'd like to pretend and play a little bit with it for a moment. Notice the increase in the r form .14 to .28, an increase in prediction of criterion variance from 2% to almost 8%, which in turn is an increase in predictive power of about 300%, not by doing any more measurements or more work, but just by changing the contents of what we already planned to do anyway.

Let's notice that the increase in the standard deviation is proportional with the increase in the correlation. And the suggestion from this data that an r=.39 could be obtained using a situational interview about 1:6 times instead of 1:10,000.

And I also want us to consider the application of the technique in a managerial setting. You can see from this data that Latham was willing to try this technique in some very unconventional settings, and his worst results occurred with his largest sample, when he was using the situational interview to hire linemen for a utility. In the same way the "worst" results I've experienced with this technique came in hiring police officers.

(However when coupled with a highly valid written test, and with the two procedures weighted roughly commensurate with their validity values, the multiple approached .9).

Why would this variation of the interview procedure yield indications of vastly superior results? In some informal remarks Jack Hunter suggested it might be because the situational interview, in miniature, is an exercise of broad analytic, problem-solving abilities. It is, in an oral format, an abilities test--and from "Alternative Predictors" we learn (if we had not learned it earlier) that nothing predicts job performance like ability. Not interest, nor college grades. Not references, nor personality tests. Not handwriting analysis, nor amount of education. Only (1) personal achievement and (2) knowledge rival ability for predicting job success.

The primary similarity between the interview proper and the situational interview is that both exercises are oral in nature. From that commonality the two rapidly depart from each other. And the differences are not primarily in format but in content.

We can say of either kind of interview that it may or may not be standardized/formatted/patterned/prescriptively documented/or "structured."

No matter what term we use, this is all the same thing. In my opinion none of it adds validity to the interview procedure, nor does it absence necessarily take away validity. "Structure" is a formalistic issue, but validity is not structurally based--it is not formalistic in nature. Validity is content-based. To improve the validity of a test, add more content to it; or improve the content otherwise, such as by making the test content more relevant to the objective, whether the objective is course learning or job mastery or whatever.

What sets apart the situational interview most from the common interview is content, their different contents. The main problem with the interview, as I see it, is that we do not know what its content is. It can have content anyone wants it to have, which is to say it has no known content. Ability tests, on the other hand, have content. Knowledge tests have known content. Assessment centers have the known content of social skills. The interview proper has no known content other than perhaps "oral communications skills," and that is typically so poorly and improperly defined as to miss a dimension as fundamental as listening skill and ability.

Situational interviews have known content as well. They derive their content from job analysis, usually the critical incident technique. They pose problems about job-performance-related situations of choice and judgment. Solutions to these problems indicate reasoning, common sense, job judgment, problem-solving ability, or whatever we want to call this factor. Whatever it is, scores on that dimension correlate moderately with other reliable, useful indicators of job performance and success.

Discussion of the content of the situational interview brings me to my specific topic today, what I have called the BOSS technique. BOSS is

simply an acronym for Behavioral Observation ScaleS. Which is to say, the evaluation criteria by which the candidate responses to the given problems are compared.

I do not think the distinguishing content of the situational interview lies in the questions, or the problems as I prefer to call them, so much as in the answers, that is, not the responses but the standardized answers found in the scales for scoring those responses. The key to this kind of interviewing is the test key itself: The Behavioral Observations ScaleS.

Where we need to start from is not what we want to say/ask the job candidates, but what we want the candidates to say and us to hear. That is, it seems to me we want to design our tests and exercises initially in terms of the information we want to get, not the probes we want to use. If the knowledge is not important, we don't want to ask about it. If the thinking is not critical, we don't want to request it.

Knowing what our answers are, what the answers should be, is more important than exactly how the questions should go. To illustrate--

You could spend a page or a phrase asking this germ of a problem, and it will all amount to about the same thing: "Your subordinate has been coming to work late the last few days..."

There are all kinds of ways of dealing and not dealing with this problem in reality, and perhaps three times as many ways of answering this question in an interview. What I suggest we need to do as test-makers is know exactly how we are going to evaluate the responses we are most likely to receive.

In BOSS scaling what we do is list and pre-evaluate all the examples of responses to which we would want to give the highest credit, and all the examples of responses to which we would want to give the lowest credit possible. Sometimes we may also list intermediate levels of responses. But the emphasis is on the Excellent level because that is the target level of ability we are trying to identify and hire. We are ultimately not interested in intermediate levels of relative ability, and are not trying to be either as exact or as certain at ranges lower than excellent. We do, however, like to anchor the lowest level in a detailed fashion so that the rating interviewers have a clear idea of what constitutes the opposite of excellence in the problem.

The judgment of the rating interviewers is confined to comparing what they have heard against the concretely and specifically detailed BOSS evaluation criteria and to discerning the proper balance of things when they have heard a mix of answers and elements of answers in a candidate's response to a given problem.

In this respect Latham's procedure is somewhat different from my own technique. Latham's scaling involves limited benchmarking of the scale points, and he makes up for it with pre-testing and with additional interviewer training. My technique involves anticipating in more detail the

likely responses and documenting them in advance so that when they occur in the interview rater error is minimized by the governance provided through the BOSS criteria.

There is considerable opportunity for further research on the interview and especially on the situational interview. But the promise and value of this technique for evaluating job candidates clearly makes it one of the superior rating procedures.

References

Huett, Dennis L. "Improving the Validity of the Interview in a Civil Service Setting," Washington, D.C., IPMA, 1976.

Hunter, J.E., & Hunter, R.F. Validity and Utility of Alternative Predictors of Job Performance. Psychological Bulletin. 1984, v.96, pp. 72-98.

Latham, Gary P., and Saari, Lise M. Do People Do What They Say? Further Studies on the Situational Interview. Journal of Applied Psychology, 1984, v.69, no. 4, pp. 569-573.

Latham, Gary P.; Saari, Lise M.; Pursell, E.D.; and Campion, M.A. The Situational Interview. Journal of Applied Psychology, 1980, v.65, pp. 422-427.

* * *

Discussant's Comments

Joel P. Wiesen, Commonwealth of Massachusetts

When I evaluate exams in court, or teach industrial psychology, I say that all exams and all methods and systems for personnel selection must pass muster in 5 areas. Let's look at these first and then consider each of the presentations with respect to these and with respect to their specific stated goals.

The five evaluative areas are:

- **Practicality** —Will people use the exam?
- **Reliability** — Are the grades replicable? (This is required for the exam to be valid.)
- **Validity** — Does the exam predict job performance? Was the development of professional caliber?
- **Utility** — What is the net monetary benefit of using the exam?
- **Legality/EEO** —Is the exam fair and are we likely to prevail if challenged in court?

Karen Duffy-Wallace and Bruce Davey have given us a wonderful example of an applied research program in a state personnel department. In 1984 they found that, despite structure in the exam, their oral panels differed in mean scores, standard deviation of the scores, KSAs emphasized and, most importantly, validity. They set about to rectify this. Their new highly structured approach maintains content validity while achieving very high structure and reliability in grading.

Roger Davis developed an oral examining approach to managerial selection using situational interviewing and 11 BOSS scales. Two types of validity evidence were presented, content and criterion related. It is always comforting to see more than one line of validity evidence, with each supporting the other. Roger is moving in the right direction: replacing the traditional interview with a more precise examining system.

Jerry Davis focused on one small part of oral examining. He developed training materials for oral raters including: a guide for oral raters, 2 videotape training films, training exercises, and a student (rater) manual. Once developed, these are relatively easy to use and the user acceptance is high. No information was presented to allow an evaluation of the reliability of the grades, nor the validity of the test nor the utility of the selection process, nor the legality. However, Pennsylvania has extensive documentation for these areas in a number of other publications.

We see here practitioners engaged in similar attempts to structure the oral exam, both in grading and in administration. There were several other presentations at the IPMAAC Conference which reported on similar efforts (one by Janet McGuire comes to mind).

We assessment specialists need to share our techniques by publishing them. These publications need to have enough detail so that others can use the techniques as written, without reinventing the many and sophisticated details of their application. Unless and until we do this, assessment will be more of a craft learned at the hand of a senior person, or reinvented many times, and less of a profession with a systematic body of knowledge. But professional journals do not publish this type of work with the needed detail. I think the IPMA Assessment Council is interested in helping assessment specialists to do just this in several ways, through detailed presentations at the annual IPMAAC Conference, through workshops at the Conference and during the year across the country, and through publications of the details of specific examining techniques. I urge IPMAAC to publish a manual of oral examining methods, including the types of methods described in this session.

* * *

How Accurate is Self-Assessment Data on Management Skill Dimensions?

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

Overview

In recent years, there has been a trend toward integrating self assessment components into selection and promotion procedures. This paper provides the results of research into how accurate an individual's self perceptions are when selection and promotion are not potentially biasing factors. Specifically, this presentation will look at the correlation between participant and assessor ratings of participant performance in several career development assessment center programs.

Each program in the study included a thorough job analysis, custom job-related exercises, and an assessor training program ranging from 10 to 16 hours prior to assessment. In each assessment center, participants were provided with detailed definitions of the performance dimensions, including ideal characteristics for each, copies of the assessor report forms (rating sheets) for each exercise within which they would participate, and a brief orientation on how to complete the rating sheets. The orientation included a description of the rating scale values and stressed that the data obtained would be valuable for determining how accurate their self perceptions were when compared with how they are viewed in the same situations by others (the trained assessors).

In addition to completing forms identical to those completed by the assessors regarding their performance in each job simulation exercise, participants completed an extensive self assessment form regarding their level of competence in the same dimension categories in general. That is, each participant was asked to describe where they used the various management skills in their everyday life (on and off the job) and then to respond to a series of questions designed to determine their self-perceived level of competence in each skill area.

This paper presents the results of comparing the assessor ratings on each dimension factor to the participant self ratings from the exercises as well as from the skills inventory. The results of this analysis should be valuable for selection specialists who have or are considering the use of self assessment data as part of their examination processes. The results should also be valuable to anyone who uses self assessment inventories as a source of information for career development programs/decisions.

The Study Design

Career development assessment centers were conducted in four different public organizations: two at the state level and two at the local government level (one county and one city). Table I summarizes the organization, assessor, participant, exercise and dimension characteristics of each of the four career development programs.

In each of these assessment centers, trained assessors evaluated participant performance in three or four job-related exercises developed specifically for the level and target occupations identified in Table I. The assessment centers were scheduled so that in each exercise, two assessors independently evaluated each participant's performance. Further, each center was scheduled so that each of the six to eight assessors evaluated each of the participants in one of the three or four exercises. Finally, the schedules ensured that each assessor evaluated some participants in each type of exercise.

TABLE I

ASSESSMENT CENTER CHARACTERISTICS

	Organization A	Organization B	Organization C	Organization D
TYPE OF ORGANIZATION:	City, Law Enforcement	State, Law Enforcement	County, Public Works	State, Health and Welfare
TARGET LEVEL OF ASSESSMENT CENTER:	Top Management	First Line Supervisor	Division Chief (Senior Civil Engineer)	First Line Supervisor
PARTICIPANT LEVEL:	Middle Management	First Line Supervisor	Asst/Assoc Civil Engineer	Journey Level Analyst
PARTICIPANT SELECTION METHOD:	Voluntary	Mandatory	Voluntary (with encouragement)	Voluntary/Lottery
NUMBER OF PARTICIPANTS:	8	45	22	24
ASSESSORS:	Inside - 1/Outside - 5	Inside - 0/Outside - 8	Inside - 0/Outside - 8	Inside - 8/Outside - 0
LENGTH OF ASSESSOR TRAINING:	Prereading Plus 8 Hours	Prereading Plus 12 hours	Prereading Plus 8 Hours	Brief Prereading Plus 16 Hours
NUMBER OF DIMENSIONS EVALUATED:	11	11	11	10
EXERCISES (See Key Below):	IB, OP, GR	OP, GR, WP	IB, OP, GR, WR	IB, GR, RP

Exercise Key

IB = Inbasket
OP = Oral Presentation
GR = Group Discussion
RP = Role Play
WR = Written Report
WP = Written Problem with a Follow-up Oral Component

Each center included integration sessions one day after the observation of eight-twelve participants in the exercises. During these sessions the assessors integrated their initial perceptions of participant performance focusing on the performance dimensions which were being assessed and revised any of their initial numerical ratings they felt on reflection were not appropriate. The assessors then developed overall recommendations to assist each participant in their individual career development efforts (i.e., no "overall score" was assigned).

Behavioral Dimensions

- Written Communication Skills*
- Oral Communication Skills
- Decision-Making Skills
- Ability to Analyze and Solve Problems
- Planning and Organization
- Awareness of Political and Social Ramifications**
- Management Control Skills
- Leadership Skills
- Interpersonal Sensitivity Skills
- Flexibility
- Composure and Self Control

*All the dimensions were defined similar to this one.
**This dimension was not assessed in Organization D.

At the beginning of each center, the participants were provided with blank assessor report forms for each exercise, identical to those which would be completed by the assessors. They were also provided with detailed definitions and a list of ideal characteristics for each of the performance factors (dimensions) being measured. Finally, the participants were oriented to the 7-point rating scale which would be used in completing the assessor report forms. This orientation stressed the importance of each participant being as objective as possible in completing their self evaluations, the goal being to see how consistent their self evaluation scores would be when arrayed next to and compared with the scores assigned by the assessors for their performance in the same situations (exercises) on the same performance dimensions.

The participants were instructed to complete the evaluations (assessor report forms) immediately after each exercise. In all programs the forms were completed before performance feedback was provided to the participants.

Overall Assessor-Self Correlations

Table II illustrates the overall dimension correlations obtained and their levels of significance for each of the four programs. These correlation coefficients were obtained by computing the relationship of all self produced average dimension scores to the assessor produced dimension averages for each participant on each dimension. The dimension averages were obtained by averaging the scores assigned for each dimension across all exercises where the dimension was measured (i.e., assessors were not asked to come to a consensus by dimension across exercises).

TABLE II

Overall Assessor-Self Correlations
By Organization for All Dimensions

Organization	N	Assessor Mean	Assessor S.D.	Participant Mean	Participant S.D.	r	t	p
A	88	3.925	.898	4.507	.695	-.011	.110	.877
B	495	3.726	.985	4.292	.779	.373	8.790	.000
C	242	3.033	1.159	3.424	.964	.391	6.582	.000
D	240	2.807	1.450	3.160	1.193	.465	8.119	.000

Comparison to a Selection Center

For a comparison between the correlations obtained from these four career development programs with the correlations obtained in an assessment center conducted for promotional purposes, 17 Police Sergeants competing in an assessment center process designed for the target level of Police Lieutenant were asked to assist with this research. The promotional assessment center utilized four custom content exercises (Group, Inbasket, Oral Presentation and a Written Problem with an oral component to present and justify the written product); eight outside assessors were used and eight common management performance dimensions were evaluated. The instructions given to the 17 Police Lieutenant candidates were as follows:

Voluntary Research Survey

Candidate ID#_____

"Please help us with a research project. The goal of this research is to determine how accurately individuals can assess and predict how they have been evaluated on management performance dimensions. On the line to the left of each dimension listed below, please indicate the score you believe you averaged in the assessment center exercises today. The rating scale runs from 0-6 and is defined on the reverse side of this sheet.

This is an anonymous survey. The individual scores on this form will not be told to anyone. We are interested in the average correlation across all participants. However, in order to compare the self-predicted scores with the actual scores received from the assessors, we need your Candidate ID# at the top of the form."

Computing assessor-self rating correlations using the same computations which produced the data illustrated in Table II resulted in the following:

N=136 cases; Assessor Mean= 3.073; SD=.999; Candidate Mean=4.452; SD=1.146; r=.068, t=.794; p=.434

Assessor and Self Correlations with Self Assessment Inventory

In addition to the assessor report forms, each participant completed a Self Assessment Skills Inventory (SASI). This inventory, which required approximately 2-2 1/2 hours to complete, asked participants to respond to a series of questions requiring the performance dimensions being evaluated. For each dimension, participants were asked to describe five activities they had been involved in recently which required use of skills related to the dimension. Using a 7-point scale with each point defined, they were asked to describe, (1) how hard it was to think of the five examples and how (2) comfortable, (3) confident, and (4) competent they felt when performing tasks which require use of skills in the dimension area. The responses to these four questions were then averaged to obtain a SASI score for each dimension.

In the assessment centers for Organizations A,B, and D, these inventories were completed during the process. Organization C required participants to complete the inventory prior to the assessment center.

Table III summarizes the overall correlations between assessor ratings from the exercises and SASI ratings for all dimensions for all participants and the correlations between self ratings from the exercises and SASI ratings for all dimensions.

TABLE III

Organization	N	ASR-SASI			SELF-SASI			SASI	
		r	t	p	r	t	p	MEAN	SD
A	88	.337	3.323	.001	.104	.976	.333	4.213	.717
B	495	.151	3.410	.001	.385	9.110	.000	4.253	.878
C	242	.024	.374	.708	.190	3.003	.003	3.623	1.129
D	230	.442	7.457	.000	.327	5.228	.000	3.773	.878

Individual Dimension Correlations

In career development programs, the usual focus is on identifying specific areas (dimensions) within which to focus individual and/or organizational career development or training efforts. To determine whether participants were able to more accurately assess their skills in some dimension areas as opposed to others, correlation coefficients were produced by dimension for the two organizations with the largest number of participants (Organizations B and D).

Table IV summarizes the correlation coefficients obtained by dimension which illustrates the Assessor-Self, Assessor-SASI and Self-SASI relationships.

Table IV

Individual Dimension Correlations
Organizations B (N=45) and D (N=24)

BEHAVIORAL DIMENSIONS	Organization B			Organization D		
	ASR/SELF	ASR/SASI	SELF/SASI	ASR/SELF	ASR/SASI	SELF/SASI
Written Communication	.412**	.324*	.584**	.237	.517*	.199
Oral Communication	.309*	.200	.349*	.459*	.375	.438*
Decision Making	.432**	.044	.359*	.411*	.439*	.355
Analyze/Solve Problems	.385**	.012	.279	.507*	.643**	.336
Planning/Organization	.406**	.240	.515**	.624**	.437*	.291
Political/Social Ramifications	.384*	.135	.476**	- - Not Assessed - -		
Management Control	.363*	.338*	.571**	.654**	.526**	.448*
Leadership	.270	.145	.289	.600**	.296	.402
Interpersonal Sensitivity	.345*	.092	.488**	.355	.517*	.277
Flexibility	.317*	.000	.345*	.558**	.435*	.357
Composure/Self Control	.392**	.157	.327*	.258	.390	.388

* $< .05$
** $< .01$

Correlations By Exercise

In recent years the personnel assessment field has acknowledged that in assessment centers we are not measuring skills by dimension. Rather, we are measuring skills by dimension within a specific situational context. For example, in a career development assessment center we do not discover or report that a person is low on interpersonal skills. Rather, in providing feedback we might say "You demonstrated only a small amount of interpersonal sensitivity in the group setting." Individuals can and do often score at different ends of the rating scale on the same dimension in two different types of exercises.

Table V presents the results of computing correlation coefficients for assessor and self ratings by exercises for Organizations B and D. The dimension scores for each exercise were totaled and averaged to obtain an exercise average as illustrated on Attachment C (the Participant Score Profile). These exercises averages were used to compute the correlations between assessor and self ratings by exercise.

TABLE V

Correlation Between Assessor and Self Average Ratings by Exercise

Exercise	N	ASSESSOR MEAN (SD)	PARTICIPANT MEAN (SD)	r	t	p
Organization B						
Role Play	44	3.742 (1.74)	4.468 (.694)	.197	1.308	.195
Group	44	3.469 (1.398)	4.209 (.793)	.635	5.334	.000
Oral Pres.	44	4.055 (.870)	4.657 (.742)	.269	1.812	.073
Written Prob.	44	3.596 (1.114)	4.605 (.826)	.404	2.866	.006
Organization D						
Role Play	24	3.052 (1.410)	3.178 (1.103)	.594	3.465	.002
Group	24	2.759 (1.663)	3.276 (1.058)	.435	2.269	.031
Inbasket	24	2.392 (1.667)	2.927 (1.412)	.631	3.824	.001

Correlations by Total Performance

When assessment centers are used for selection and promotion purposes, the participant's total performance in the process is used as an indicator of potential for success at the target level. Using total of dimension scores as total or overall performance, one final correlation coefficient was computed for both Organizations A and B using assessor and self data. Table VI provides these data.

TABLE VI

Total of Dimension Scores Assessor-Self Correlations

Organization B (N=44)			Organization D (N=24)		
r	t	p	r	t	p
.513	3.874	.000	.595	3.480	.002

Brief Summary - What Do These Data Suggest?

These data suggest that there is a positive relationship in career development oriented assessment center programs between self ratings on self assessment inventories and on the exercises when compared to the ratings assigned by experienced managers working as trained assessors. This positive relationship is not a strong positive relationship. In fact, inspection of the raw data for all four assessment centers produces cases of extreme over and under-rating by self raters when compared to the assessor ratings on the same dimensions.

Overall, there seems to be sufficiently high correlations for the majority of participants to see and understand the perspective of the assessors when provided with the narrative descriptions which are provided with the performance scores in feedback. On the other hand, if one assumes that the trained assessors with more management experience are producing more accurate evaluations than the self raters, then some serious questions must be raised regarding the use of self assessments as the sole source of information upon which to base career development programs, as is quite often done. Even more questionable, would be the use of self assessment as a weighted factor in a promotion or selection process.

These data also provide further support for the often replicated (in recent years) finding that we are not measuring eight to twelve discrete dimensions across a number of exercises as much as we are measuring overall performance within exercise situations. This is supported by the higher correlations obtained when correlating assessor and self ratings by exercise. It appears that despite requiring raters to provide narrative comments to articulate, explain and justify the scores they assign by dimension, the situational context or overall problem being dealt with is the major

determinant of a participant's scores. The message here for selection specialists and career development specialists alike is that as much importance should be put on the exercises developed as on the specific dimensions which are measured. In other words, we should not develop or use off-the-shelf exercises which we believe are going to give a good measure of the dimensions determined to be important for success on the job, unless they also are fairly accurate simulations of the most important and frequently performed tasks or activities one would have to perform in the target job/classification.

This study found that the correlations between self ratings and assessor ratings are higher in the career development programs than in the (control) assessment center being used as the ranking component in a promotional examination process. Further, the mean self ratings assigned in the career development programs are lower and closer to the mean ratings assigned by the assessors. These trends are further supported by the results obtained in Organization A where in addition to individual career development the other major stated objective of the program was succession planning.

Limitations of the Data

The most important limitations of the conclusions reached in this paper are the small sample sizes. The trends identified are important if they continue to emerge through further replication.

Acknowledgement

The author would like to express sincere appreciation to Phil Carlin of Real Time Technologies based in Tucson, Arizona for his assistance in computing all data contained in this paper.

* * *

An Examination of Clerical Selection Procedures

Terry S. McKinney, Employment Services Division, City of Phoenix, AZ

INTRODUCTION

The successful recruitment and the selection of entry-level clerical employees is critical to the efficiency of any organization. This is especially true with the City of Phoenix. The citizen-taxpayer's first contact whether in person or on the phone, with most departments, is normally with a clerical support employee.

In recent years, there has been increasing concern over the quality of individuals entering City service and/or the adequacy of the selection tools used by the City of Phoenix Personnel Department. In an attempt to address the many concerns, the Personnel Department developed a questionnaire to survey the opinions of the users of eligible lists provided for entry-level clerical positions. Approximately 150 questionnaires were sent out to various City departments. The survey technique was to utilize the Personnel Officer as a contact point in those departments that had Personnel Officers. For other departments, the membership list of SHARE (Secretaries Helping Administrators Realize Expectations) was utilized. Fifty-five usable questionnaires were returned by the deadline.

Editors Note: A more recent, but similar study entitled "A Survey of Foreman Selection Procedures" was done by the Personnel Department of the City of Phoenix. A questionnaire was developed to survey the users of eligible lists provided for entry-level field supervisory positions, defined as positions in which one directs a crew or a group of Unit 1 or Unit 2 employees. A number of recommendations were made, most notably that additional improvements in the selection system need to be a continuing priority of the City's personnel department. Due to limited space, this study is not to be included in the Proceedings.

This report discusses the findings in the survey itself and identifies a number of recommendations to improve the quality of the City's entry-level selection procedures.

Part One

The first item on the survey dealt with how frequently respondents utilized our eligible lists. The data indicated that the average individual respondent had used our eligible list an average of 2.2 times in the last year. This relatively low rate of using the eligible list indicates that most of the respondents were basing their views on a fairly small sample. It is interesting to note that approximately 9% of the respondents had not utilized our eligible list in the past year. Thirty percent had used it once and 28% had used it twice with 18% using it three times. One individual respondent had utilized the list 14 times in the past year.

One concern of the City's personnel department is always the timeliness of the response to operating departments in providing an eligible list. The survey results indicated that 73% of the respondents received an eligible list within one week of their request. In general, a week's turnaround time to obtain an eligible list can be considered a timely response.

Departments have an opportunity to visit the Personnel Department to review the hard copy of the application. The survey indicated that approximately 68% of the respondents did take advantage of this opportunity to review the application. Thirty-two percent did not.

For those individuals that reviewed the eligible list (N=36), data was collected as to the major elements looked for. It is significant that 71% of the respondents indicated they looked for the level of experience of the applicants including such things as complexity of jobs held, etc. Thirty-four percent of the respondents looked at job history, length of employment, reason for leaving, etc. Ten percent reviewed the applicant's training and experience. These are all relevant and job related factors to review in determining who to interview off the eligible list.

Only 1 respondent indicated that he/she looked at the test score. This low rate in terms of examining test scores would indicate that hiring officials either find that our tests are relatively meaningless or lack an awareness as to the utility of test scores and the formatting of our eligible list. Additionally, it is surprising and disappointing to find that 9% of the respondents indicated they attempted to identify personality traits from the application. Inferences were this made about such constructs as adaptability, etc. This is probably not an appropriate conclusion or inference to draw from the application.

Results show that 55% of the respondents felt that many or most of the individuals were no longer available for work when contacted. This is an alarmingly high rate and indicates that our eligible lists are not up to date.

The respondents indicated that they interviewed an average of 7 applicants to fill a particular vacancy. The number of individuals interviewed ranged from a low of 3 to a high of 23 per vacancy.

The next section of the survey asked the respondents to indicate the relative skill or quality of the individuals they have interviewed off the eligible list in a number of different categories. Some of these results indicate significant areas for training needs while others indicate areas where our testing might be improved. Somewhat disappointing to the Personnel Department was the fact in many areas none of the respondents felt that our applicants were excellent and in only two areas, did more people rate the applicants as excellent than did unacceptable. These were telephone skills and in ability to operate office equipment (generally defined as equipment such as photo copiers, etc.).

For a number of years now, due to administrative concerns in terms of cost and scheduling, the City's Personnel Department has not conducted typing tests for our entry-level positions. Seventy-nine percent of the respondents indicated they currently administer their own typing test for these entry-level positions. Seventy-six percent felt that the Personnel Department should administer a typing test. This clearly indicates that the hiring officials surveyed view typing skills as a very important factor and while they are currently administering their own test, would prefer that this be done by the City's Personnel Department.

Respondents were asked if the City's current procedures were providing good quality candidates. Forty-six percent of the respondents were generally satisfied while 54% were not. This again clearly indicates that some modifications to the current process are necessary in order to provide and to meet the needs of the operating departments.

The survey indicated that approximately 8% of the respondents felt that applicants are better today than they were 3 years ago while 35% have thought there has been a decrease in quality of applicants.

Part Two

Some of the recommendations that follow are based on the survey. Others are based on discussions that have been conducted with hiring officials, members of the SHARE, and Personnel Department staff. It is recognized that many of these recommendations are beyond the scope of the Personnel Department or any individual department to implement. Due to the fact that a large number of respondents felt that the availability of applicants was still a problem, it is suggested that the Personnel Department explore the possibility if increasing the frequency of testing to three or four times per year.

In reference to advertising entry-level clerical positions, it is recommended that a display ad be used and that greater emphasis be given to the benefits of working for the City in terms of the career opportunities for those that join us at the entry-level clerical position. Since many positions with the City are limited to a promotional basis, it is to the City's benefit to attract individuals at the entry-level who have the skills and the ambition to move upward in the organization.

Since a large number of respondents felt that our current eligibles are deficient in a number of significant skill areas, Personnel should explore the possibility of direct recruiting through the clerical blocks at some of the schools and/or business colleges. Applicants from these areas, while they may have limited "hands on" experience, could be expected to have very high technical skills in the area of typing, etc. Perhaps the use of a working title may improve recruitment as such titles may be more attractive to potential applicants than do the traditional titles.

It is further suggested that a supplemental self certification be added to the application process. While self certification of typing skills is far less accurate than a skill test, this would at least add some information as to speed and error rate.

Those areas of the survey that had a higher rate of "unacceptable" such as proofreading ability, grammar, vocabulary, and punctuation should be reviewed by Personnel in terms of testing. The amount of the test related to these areas should be increased.

Those areas of the survey that had a low rate of "good" or "excellent" may be priority areas when training of current employees is needed. It is suggested that SHARE and Value Management examine this possibility.

Another option would be a suggestion to utilize the City's "trainee" or "noncompetitive promotional" procedure. Individuals would be hired into the entry-level or trainee class. Upon completion of a competency based training program, individuals could be promoted to target journey level class.

If the planned follow-up research is favorable with entry-level blue collar classifications, it is suggested that the Worker Opinion Questionnaire (WOQ) type tool be modified and "tried out" as part of the selection process for entry-level clerical positions.

It is clear from the survey that some users of our eligible lists lack correct information as to how names are ordered on the list and also on information available (and its proper use) on the application form. It is recommended that the placement section of Personnel work with the Personnel Officers and the EEO function to prepare the needed educational material.

* * *

A Program for Certification of the Competency of Personnel Professionals

William Maier, Colorado Personnel Department, Denver, Colorado

Background

The State of Colorado is one of two states with Constitutional requirements for a State personnel system. Besides making Personnel one of the twenty major Departments of the State, the Constitution mandates that "Appointments and promotions to offices and employments in the personnel system of the state shall be made according to merit and fitness, to be ascertained by competitive tests of competence." This constitutional requirement for "competitive tests of competence" provides the basis for a diversified testing program concerned with test quality and validity.

Colorado's testing program requires by rule and procedure that each newly developed exam be based on a job analysis. Many types of exams are typically used in compensatory and non-compensatory examination plans. These include three types of ratings of training and experience, structured oral boards, role plays, written essays, written multiple choice exams, assessment centers, physical agility exams and other types of performance exams.

Part of the philosophy of decentralization to operating agencies was a mandate that only agencies who have certified personnelists may be decentralized for the areas in which certification exists. Personnel certification of individuals, post audit of operations and appeals allow us to manage the decentralized system.

The personnel certification program is the newest of these three methods of managing the technical competence in a decentralized environment. It was implemented in the beginning of the 1986 calendar year. So far we have developed the training courses and the written multiple choice competency exams for 5 areas. These include selection, classification, performance appraisal, affirmative action and personnel rules. Selection and classification certification are only at the first level this year. The second level will be developed for implementation next year.

Levels of Certification

The level concept of certification was designed to tailor the amount of training and testing to the needs of each agency and individual. Small agencies which typically do limited testing may require only first level certification which allows the person to do the minimum set of activities necessary for simple test development and administration. A large agency which uses a number of sophisticated devices such as multiple choice examinations or assessment centers may need a person certified at a level III in selection.

The first level is characterized as "cookbook" the second as "working level" and the third as "advanced professional." Permitted activities range from developing examination plans and using written multiple choice examinations at the first level to doing criterion-related validity studies and developing written multiple choice examination at the third level.

Examinations

The examinations for the first level of selection functions are all multiple choice and based on a content domain for the module entitled "Elementary Principles of Selection and Job Analysis."

The content domain for the written tests was specifically defined using a reading list and is divided into seven modules:

1) Elementary Principles of Selection and Job Analysis
2) Examination Planning
3) Use of Written Objective Tests
4) Development of Oral Board Examinations
5) Development of Checklist Ratings of Training and Experience
6) Examination Administration
7) Legal and Professional Standards

Individuals may take all the tests at one time and then take training for those which they fail or they may take the training followed by the test. Other areas such as classification elected to give a single test and training session. The seven tests for selection contain between 50 and 85 items each. The classification test is 180 items in length and the other areas run from 50 to 100 items each.

To be certified in the first level of selection, individuals must pass all seven selection modules. Pass points for each of the certification tests were set using Nedelski's method. As many of you know the Nedelski method requires subject matter experts to judge whether or not a minimally competent person might not be able to eliminate a distractor. The probability of a minimally competent individual getting the item correct is equal to one divided by the number of plausible answers, i.e., the correct answer plus the number which the minimally competent individual could not eliminate. The sum of these will be the pass point. Of the 344 tests taken so far in all areas, 280 people passed for a pass rate of 81%.

Although we recognized that knowledge of the area was necessary but not sufficient to demonstrate competence, we decided not to do a performance exam the first year because of the large number of working personnelists who must be certified. A performance exam would be administered to each individual rather than on an assembled basis. We intend to form a professional standards certifications. This committee will decide whether or not to go to a performance exam next year.

Training

The training program for selection level I is divided into the same seven modules as the tests and required 2 to 4 sessions of four hours each. We ran one session each week in the hope that spreading the course out will allow people to devote more time to learning.

Problems and Results

We expected a good deal of resistance from people who must be certified. We received some complaints and foot dragging, but in general there was much less resistance than originally anticipated. One of the reasons seem to be that the most competent people in each field were generally supportive of the idea. They had seen some real problems as a result of the loss of technical competence. Whey they took the tests and had no trouble passing, their support for the program grew.

We do not yet know what will happen if a decentralized agency does not have a certified individual at the end of this year. At that point, they will not be able to sign-off on the creation of eligible lists or job audits. We are hoping all decentralized agencies will have a certified professional.

All but one of twenty agencies decentralized in examinations have participated in the first testing and training. Six agencies already have at least one individual certified in selection.

The program appears to be increasing the quality of the work done in selection. In the future, we will be able to compare the quality of tests developed before implementation of the certification program with those developed after its implementation through the quality review part of our post audit program. We currently have no objective measure of a change in quality, but we are getting more questions about developing quality tests and there is more concern expressed about how good a test is. We believe this program is increasing a sense of professionalism and is the corner stone of the management of a decentralized personnel system.

Note: Additional materials related to this article are not included due to space limitations may be available from the author.

* * *

SUBJECT INDEX

AUTHOR INDEX