

## DOCUMENT RESUME

ED 337 485

TM 017 306

TITLE Proceedings of the 1990 IPMAAC Conference on Personnel Assessment (14th, San Diego, California, June 24-28, 1990).

INSTITUTION International Personnel Management Association, Washington, DC.

PUB DATE Jun 90

NOTE 549p.

PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF02/PC22 Plus Postage.

DESCRIPTORS Assessment Centers (Personnel); Computer Assisted Testing; Job Analysis; \*Job Performance; \*Occupational Tests; Personality Assessment; \*Personnel Evaluation; Personnel Management; Personnel Selection; \*Public Sector; Test Use

IDENTIFIERS International Personnel Management Association

## ABSTRACT

Fifty-seven papers presented at the annual meeting of the International Personnel Management Association Assessment Council (IPMAAC) in 1990 are provided. Selected topics include: using the cloze technique for reading skills assessment; examining assessment techniques; job analysis; alternate strategies for assessing writing skills; assessment of workforce quality and employability skills; the use of job analysis in promoting workplace justice; advances in multiple-choice item writing and review; techniques to select workforce 2000 and its leaders; developments in personality measurement; the validation of R. Hogan's Prospective Employee Potential Inventory on school bus drivers; using video technology in the selection process; directions for the manager of tomorrow's assessment function; innovations in peace officer selection; recent innovations in public sector assessment; Navy research on advanced technologies for selection and training; effective assessment practice for the next century; finding and assessing selection instruments and consultants; structured interviewing; electronic document storage--a case study; evaluating writing skills; using today's techniques and technologies to prepare for assessing the workforce in the year 2000; validation strategies; decentralizing an automated performance evaluation system; assessing clerical skills impacted by office automation; work sample based selection for assembly personnel meeting the demands of a Japanese management climate; selection criteria for the 1990s--implications for job analysis; progress in assessment center methodology; use of non-traditional training and experience ratings; what bio-data predict; personnel selection that meets the evolving legal requirements; the effects of candidate orientation, candidate training, coaching and management/supervisory training on assessment center performance; assessing physical ability; job qualification linkage system; and job analysis approaches. (SLD)

\*\*\*\*\*

\* Reproductions supplied by EDRS are the best that can be made \*

\* from the original document. \*

\*\*\*\*\*

# IPMA Assessment Council

ED337485

## Proceedings of the 1990 IPMAAC Conference on Personnel Assessment

U.S. DEPARTMENT OF EDUCATION  
Office of Educational Research and Improvement  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS  
MATERIAL HAS BEEN GRANTED BY

MARIANNE ERNESTO

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)."

June 24-28, 1990  
San Diego, California

Tm017306

**TABLE OF CONTENTS**  
**(First Authors Only Listed)**

PAGE

**IPMAAC KEYNOTE ADDRESS**

Looking Ahead to the Year 2000: Milton D. Hakel.....	1
Looking Back on the Year 2000--What We Should Have Done: Milton D. Hakel.....	3

**USING THE CLOZE TECHNIQUE FOR READING SKILLS ASSESSMENT**

An Introduction to the Cloze Technique of Reading Skills Assessment: Michael J. Dollard.....	9
---	---

**EXAMINING ASSESSMENT TECHNIQUES**

Employee Referral Programs: Do Successful Employees Refer Better Applicants Than Unsuccessful Employees?: Michael G. Aamodt.....	24
---	----

**JOB ANALYSIS: APPLICATIONS AND INNOVATIONS**

Job Analysis: Applications and Innovations: Susan R. Lozada-Larsen.....	28
--	----

**ALTERNATE STRATEGIES FOR ASSESSING WRITING SKILLS**

Training Assessors to Produce High Interrater Reliability In Evaluating Complex Writing Samples: Jade Kuan Hoffman.....	40
--	----

**ASSESSMENT OF WORKFORCE QUALITY AND EMPLOYABILITY SKILLS**

The "Nestor Factor:" Measuring Quality in the Science and Engineering Workforce: Jeanne Carney.....	49
OPM's Quality Assessment Program: Jay A. Gandy.....	59
Michigan's Approach to Assessing Employability Skills: Paul M. Stemmer, Jr.....	61

**THE USE OF JOB ANALYSIS IN PROMOTING  
WORKPLACE JUSTICE**

Job Analysis, Performance Appraisal, and Organizational Justice: Thomas J. Atchison.....	72
---	----

**ADVANCES IN MULTIPLE CHOICE ITEM WRITING AND REVIEW**

Test Item design and Evaluation: Thomas M. Haladyna.....85

**TECHNIQUES TO SELECT WORKFORCE 2000 AND ITS LEADERS**

Leaders for Workforce 2000: Innovative Strategies for Meeting  
Selection Needs: Jamie J. Carlyle.....100

Multiple Hurdle Selection of Transit Supervisors:  
John C. Barber.....112

A Work Sample Performance Test that Truly Recreates the Job:  
Robyn Wachtel.....130

**NEW DEVELOPMENTS IN PERSONALITY MEASUREMENT**

New Developments in Personality Measurement: Robert Hogan.....140

**THE VALIDTION OF ROBERT HOGAN'S PROSPECTIVE EMPLOYEE  
POTENTIAL INVENTORY ON SCHOOL BUS DRIVERS**

The Prospective Employee Potential Inventory: A Validation Study  
With School Bus Driver: Thung-Rung Lin.....152

**UTILIZING VIDEO IN THE SELECTION PROCFESS**

Job Simulation Training & Feedback Sessions: The Good, The Bad,  
& The Ugly: Jeff Prewitt.....161

**DIRECTIONS FOR THE MANAGER OF  
TOMORROW'S ASSESSMENT FUNCTION**

A Shopper's Guide to Personnel Assessment Professionals:  
Jeffrey P. Feuquay.....166

**INNOVATIONS IN PEACE OFFICER SELECTION**

Innovations in Peace Officer Selection: The California Highway  
Patrol Experience: Bob Giannoni.....174



# RECENT INNOVATIONS IN PUBLIC SECTOR ASSESSMENT

Recent Innovations In Public Sector Assessment:	
Charles F. Sproule.....	180
Selected Innovations in a State Merit System: Paul D. Kaiser.....	185
Innovations in Public Sector Assessment Centers: Dennis Joiner...	192
OPM's Major Program Initiatives in Personnel Research and Development: Jay A. Gandy.....	194
A Statutory Authorization for Selection Experimentation:	
Julie Vikmanis.....	200

## NAVY RESEARCH ON ADVANCED TECHNOLOGIES FOR SELECTION AND TRAINING

Improved Scoring for Personnel Tests: J. Bradford Sympson.....	210
Biographical Data: The Past Predicts the Future:	
Mary A. Quenette.....	218
Computer Based Instruction Technology: Douglas Wetzel.....	227
Analysis of Human Brain Electrical Activity: Towards Real-Time Prediction of Human Performance: Leonard J. Trejo.....	237

## EFFECTIVE ASSESSMENT PRACTICES FOR THE NEXT CENTURY

Learnings from an Affirmative Action Effort: David Lopez-Lee.....	245
---	-----

## FINDING AND ASSESSING SELECTION INSTRUMENTS AND CONSULTANTS

Role of the Buros Institute of Mental Measurements as an Information Provider to Personnel Selection Specialists:	
Barbara S. Flake.....	264
Hiring and Monitoring Consultants: Vicki Packman.....	274

## STRUCTURED INTERVIEWING: THEORY, RESEARCH AND PRACTICE

The Applicability of the Situational Interview and the Patterned Behavior Description Interview for Entry and Managerial Level Jobs: Heidi H. Hrowal.....	278
---	-----

# **ELECTRONIC DOCUMENT STORAGE: A CASE STUDY**

Electronic Document Storage A Case Study: Ted Daraney.....290

## **EVALUATING WRITING SKILLS**

An Investigation into the Interchangeability of Essay and Multiple Choice Tests as Measures of Writing Ability:  
Anne Forinash Friend.....296

## **USING TODAY'S TECHNIQUES AND TECHNOLOGIES TO PREPARE FOR ASSESSING WORKFORCE 2000**

Using Today's Techniques and Technologies To Prepare For Assessing Workforce 2000: Kay Barrow.....313

Automation of Tennessee's Employment System: Robert Perry.....323

## **VALIDATION STRATEGIES**

The High-Low Predictive Validity Design: High Power With Small N: Joel P. Wiesen.....330

## **POWER TO THE USERS: DECENTRALIZING AN AUTOMATED PERFORMANCE EVALUATION SYSTEM AND NOT REGRETTING IT**

"Power to the User -- Decentralizing an Automated Evaluation System and Not Regretting It": Grant Gilfeather.....344

## **ASSESSING CLERICAL SKILLS IMPACTED BY OFFICE AUTOMATION**

Assessing The Impact of Office Automation Technology on Secretarial and Clerical Jobs: Marianne Bays.....349

Performance on a PC-Based Typing Test (R.D. Craig) Versus Performance on a Traditional Typing Test: Barbara E. Leighton....360

## **WORK SAMPLE BASED SELECTION FOR ASSEMBLY PERSONNEL MEETING THE DEMANDS OF A JAPANESE MANAGEMENT CLIMATE**

Job Analysis Across Two Cultures -- U.S. and Japan: Collection Accurate Data Without the Use of Job Incumbents:  
Kevin G. Love.....372

# SELECTION CRITERIA FOR THE '90'S: IMPLICATIONS FOR JOB ANALYSIS

Findings and Recommendations from a Multi-Purpose Job Analysis  
of First-Level Supervisory Classes: Donna L. Denning.....378

## PROGRESS IN ASSESSMENT CENTER METHODOLOGY

A Career Development Assessment Center for Court Manager:  
Patrick T. Maher.....390

San Diego County Management Academy: A Program that Succeeded:  
Del Boerner.....394

A Follow-up to the 1989 Criterion Related Validity Study  
Conducted to Select Lieutenants in the Palm Beach County Fire  
Department: Linsey Craig.....397

## USE OF NON-TRADITIONAL T&E RATINGS

Use of Non-Traditional Training and Experience Rating:  
Barbara Kervi.....412

## WHAT DOES BIODATA PREDICT?

Can Biodata Predict Performance?: Herbert George Baker.....424

Can Biodata Predict Personality?: Terry W. Mitchell.....430

## PERSONNEL SELECTION WHICH MEETS THE EVOLVING LEGAL REQUIREMENTS

Recent Trends in Legal Requirements and Personnel Selection In  
The Public Sector: Gerald V.Barrett.....434

Pressures to Use Unwarranted Procedures to Reduce Adverse  
Impact in Public Sector Personnel Selection:  
Ralph A. Alexander.....450

Information Processing Approaches to Test Development and  
Construction as Evidence for Test Validity: Dennis Doverspike....459

**WHAT ARE THE EFFECTS OF CANDIDATE ORIENTATION, CANDIDATE  
TRAINING, COACHING AND MANAGEMENT/SUPERVISORY TRAINING  
ON ASSESSMENT CENTER PERFORMANCE**

Candidate Preparation for Assessment Centers:	
Janine P. DuMontelle.....	469

**ASSESSING PHYSICAL ABILITY**

Innovative Methods for Cut Scoring Determination and Injury Prevention that Improve the Bottom Line (\$):	
Deborah L. Gebhardt.....	472
Ergonomic Principles and the Development of Physical Ability Standards: Oscar Spurlin.....	478

**JOB QUALIFICATIONS LINKAGE SYSTEM**

Job Qualification Linkage Systems: S. Morton McPhail.....	486
---	-----

**JOB ANALYSIS APPROACHES**

The Hierarchical Job Analysis: A Structured Approach to the Job Analysis Interview: David E. Smith.....	499
Toward a Generic Job Analysis System: Jai Ghorpade.....	513
A Comparison of Job Information Descriptors for Classifying Jobs for Selection Purposes: Nelson Adrian.....	525

# Looking Ahead to the Year 2000

Milton D. Hakel

University of Houston and  
Organizational Research & Development, Inc.

Keynote Address, 14th Annual Conference, International Personnel Management Association Assessment Council, San Diego, California, June 25, 1990.

First, the art of making predictions is briefly examined together with a review of their accuracy. Then, following a presentation on the changing demographics of the work force, we look back at the year 2000 and what we should have done to get ready for it. The lessons: Keep our core values before us and speak the truth clearly, advance our science and practice of assessing and developing talent, and become more aware of our global interdependence.

My task for the next few minutes is to look ahead to the year 2000, to try to anticipate the future, so that we may prepare for it effectively.

In preparing for this task, I was impressed by the spate of year end and decade end reviews that appeared in the press and on TV. I also couldn't help but notice the large number of conferences in which the year 2000 plays the thematic role, for example, this one (obviously), the assessment center congress, and several conferences on technological innovation. Just last Friday, National Public Radio announced a series of features on America 2000, and invited listeners to send in suggestions.

All this predicting made me wonder about predictions made in the past, and how accurate they have been. I didn't have to look far.

Here's a chance for you to try your hand, in this case at postdiction. Turn your watches back, not an hour, but a decade, and forecast several items for 1990. The categories come from a contest sponsored by *Forbes* magazine in 1980. There were 35 predictions to be made in the contest, and the winner collected \$10,000. There were 660 entries.

## Forecasting for 1990

	1980	1990
Price of gold	\$525	\$399
N of Malls	19,200	34,000
N of Computers	300K	52M
N of Lawyers	464,851	725,574
N of NFL Teams	28	28
National Debt	\$845B	\$3,000B

*Forbes*, 1990

*Forbes* has announced a new contest, with a \$100,000 prize, for predicting what things will be like in the year 2000.

You know how touchy computers are--transpose a couple of digits and you get something other than what you started out to get. I did a computer search to get ready for this talk. Well, 1980 transposes easily to 1890, and people 100 years ago were making predictions too.

## PREDICTIONS FROM 1890

- All the trees will be gone
  - Growing demand for heating and cooking wood
  - High cost of coal and oil
- Quiet streets
  - No crash of horses hoofs, no rumble of steel wheels, due to automobiles
- Bullet trains
  - End of steam locomotives, trend to electric
  - Albertson's magnetic train, from New York to San Francisco, 3000 miles in 10 hours

*The Futurist*, 1990

I also found an interesting comment on economics:

"It is in this world market that the manufacturers and capitalists of the United States, as well as those of England and Continental Europe, must hereafter compete with one another. Powerful influences have swept away the natural barriers to competition by making it possible to transfer goods at small cost from the place of production to the remote corners of the world" (Conlon, 1890). These comments are still true.

How does one succeed at making predictions? Well, straight line extrapolation is one method, which obviously doesn't work. Incomes and stocks occasionally go down, as well as up. 100 years ago, 95% of the labor force was engaged in food production and food distribution. Experts predicted that when the farms went, the world would be put out of work. Most of us are still working.

Inspection of trends is another method. It is better, but not flawless. For example, here is a prediction of mine from 20 years ago, a symposium proposal that was rejected by a program committee:

"Assessment centers are sweeping the country. It seems reasonable to guess that centers will show the same historical cycle as flannel boards and sensitivity training. Flannel boards led to better presentation of information. Sensitivity training fostered greater concern on the part of managers for the importance of interpersonal interaction. Assessment centers will probably leave greater concern for the measurement of performance. We hope there will be some useful residue when the boom has run its course."

Assessment centers are still vital and expanding. This is the IPMA Assessment Council. So, you are now forewarned about my forecasting prowess.

The best approach to prediction is model building, using data from the past to create regression equations. The science of personnel selection is a fine example of this approach, and it may be one of the most successful examples. Meta-analysis results show that selection tests are effective prediction devices.

Model building has been applied in many areas, and is the basis for the great concern about Work Force 2000. It starts with things we know, such as population demographics. John Maynard Keynes observed that the great events of history are often due to slow changes in demography, hardly noticed at the time.

Such may still be the case, for while the *Work Force 2000* report has attracted a great deal of attention, it will be several years before we know whether the attention was sufficient. To make sure that you know about it, the program committee invited Mr. Irv Margol to address us on the changing demography of the work force.

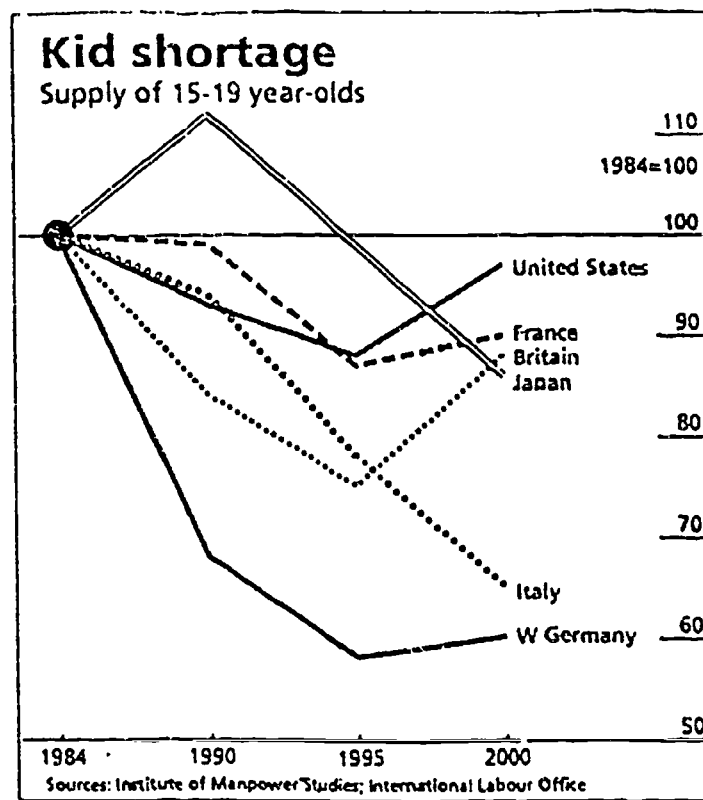
At this point, Irving Margol, Executive Vice President of Security Pacific Corporation, presented an address entitled "The Changing Demography of the Work Force."

In searching the literature, I found a few more predictions for the 1990s, to add to those that Irv presented, and here they are:

#### PREDICTIONS FOR THE 1990'S

- Married couples working together
  - "Copreneurs"
  - Starting own businesses
  - Hired as teams
- Retirement may become a thing of the past
  - Senior apprentices
  - Consultants, part-timers
- Telecommuting, flex-place, and flex-time
  - Single parent households
  - Stress of urban commuting
- Growth of service sector
  - 95% of all U.S. jobs

- Information management systems
  - 90% of service workers using CRTs by '95
  - Increasing literacy requirements
  - Reduction in layers of management
- Rate of promotion to top management
  - 1987: 1 in 20
  - 2001: 1 in 50



Only hindsight is certain, so let's once again do a little time travel, this time to the future. Set your watches ahead, not just a day, but a decade, or just a bit more, so that we can be looking back on the year 2000. Close your eyes, fasten your seatbelts, click your heels together, it's back to the future. We're not in Kansas any more.

25th Annual Conference

IPMA Assessment Council

June 25, 2001



# Looking Back on the Year 2000--What We Should Have Done

Milton D. HakeI

University of Houston and  
Organizational Research & Development, Inc.

Welcome to the 25th annual IPMA Assessment Council conference. My task today is to talk with you about what we should have done, back in 1990, to get ready for what we knew was coming.

Let me begin with an affirmation: It is a powerful statement and it has set the course for this country for the past two hundred and twenty-five years: *"We hold these truths to be self-evident, that all men are created equal; that they are endowed by their creator with certain unalienable rights; that among these are life, liberty, and the pursuit of happiness."*

No other single statement expresses so clearly the dominant values of American life. It is an affirmation that permeates our society. Every child learns it in school. These basic principles -- equality, life, liberty, pursuit of happiness -- are known throughout the world.

This affirmation from the Declaration of Independence finds expression in the Constitution of the United States. The due process and equal protection clauses of the Constitution provide the basis for the commonly accepted idea that all persons similarly situated be treated alike, both in the privileges conferred and in the liabilities imposed.

We need to keep this core value before us.

As we look back at the late 20th century, I want first to briefly review the history of the Civil Rights Act of 1964, and then offer several observations about what we should have done following its passage. We needed to speak the truth clearly and plainly. We needed to advance our science and our practice of assessing and developing talent. And we needed to become more aware of our interdependence.

First, some history. Hubert Humphrey led the floor fight in the Senate when the Civil Rights Act of 1964 was being debated. He reports that upon his first meeting with President Johnson, Johnson launched into one of his traditional speeches about liberals: "You bomb throwers make good speeches. You have big hearts. You believe in what you say you stand for, but you're never on the job when you need to be there. You spread yourselves too thin making speeches to the faithful" (Humphrey, 1976, p. 274). He said that liberals never had really worked to understand the Senate rules and how to use them;

that they were never organized effectively and would therefore go down to defeat. Humphrey wrote later that he would have been outraged if Johnson hadn't been basically right and historically accurate.

The liberals got organized, and the bill passed. The legislative history of Title VII and subsequent judicial decisions made it abundantly clear that merit was to be the basis for making selection decisions. The net effect of living with Title VII has been to reinforce the use of appropriate and improved procedures for making personnel decisions. This is clearest in the concept of job relatedness which was codified in *Griggs*. It is a simple and compelling notion that people should be selected based on their ability to do the job. It is fully consistent with our ideal of fairness, and it is, of course, the embodiment of the principle of merit.

What has not been clear in the intervening years, however, is the operational means by which merit should be measured. Who requires "equal protection"? Who is "similarly situated"?

On the surface, selection based on standardized test scores seems to satisfy the concept of fairness. Tests are color blind, and all test takers are treated alike, without regard to race, sex, religion, or other irrelevant considerations. The problem, however, is that test scores, by their specificity, narrowness, and appearance of precision, may inhibit the introduction of other considerations relevant to the employment decision. Test scores typically reflect only one or a few dimensions of individual differences. The "whole person" is not measured by a single test score, and yet the "whole person" is hired for the job. What we lack are reliable and accurate measures of other employee characteristics that job analysis may show to be important.

In a 1982 report, the Committee on Ability Testing of the National Academy of Sciences noted, "The diminished prospects of the average American give the demand about testing an especially sharp edge. Because tests are visible instruments of the process of allocating economic opportunity, tests are seen as creating winners and losers. What is not as readily appreciated, perhaps, is the inevitability of making

choices: whether by tests or some other mechanism, selection must take place."

Then, along came the Civil Rights Act of 1990, and you know how that came out.

Concerning the Civil Rights Act of 1964, Hubert Humphrey wrote, "A great barrier began to swing aside on the day when the civil rights bill passed -- a barrier that for generations had been damning back tremendous intellectual resources, incalculable energy and vitality lost to the American nation. I say 'began.' For if the door is being unlocked, if the door is swinging open, it is only just ajar. There will be no miracle wrought over night. Rather, then will come the real test of the maturity of our people" (Humphrey, 1976, p. 105).

Well, we are a long way into the "real test of the maturity of our people." How are we doing?

#### WOMEN ENTER THE WORK FORCE

	1970	1990 (est.)
New M.D.s	8.4%	31%
New Lawyers	6.4%	39%
New Dentists	1.0%	23%
Female Percent of Work Force	38%	45%

*The New Republic*, 1989

Back in '90, it was clear that women had made much progress, though there was still much to be made. Entry to the professions showed the most progress, but sexism was still a big problem.

#### EQUAL OPPORTUNITY?

	Whites	Blacks
Median Family Income	100	56
Proportion Below Poverty Line	10%	32%
Children Below Poverty Line	15%	45%

*The Economist*, 1990

But it was also clear that for African-Americans, we were failing. America was failing. Racial prejudice remained strong.

Enough of history. Let me now turn to what we should have done.

In looking back, it is clear that first and foremost, we should have spoken the truth.

Vaclav Havel, first a dissident playwright and then president of Czechoslovakia, wrote about our global crisis, which embraces Western civilization, too -- the people of London and New York as well as Prague

and Moscow. He called us to reaffirm the old-fashioned values that represent our truly human qualities: to speak difficult truths without fear, to adopt strong principles and stick to them, to expose all forms of hypocritical cant, to resist encroaching tyrannies. "Even the toughest truth expressed publicly...suddenly becomes liberating" (Havel, 1990).

In this context, let me remind you of the words of Professor Shelby Steele, a young African-American who spoke clearly and courageously about affirmative action in 1990:

One of the most troubling effects of racial preferences for blacks is a kind of demoralization. Under affirmative action, the quality that earns us preferential treatment is an implied inferiority.

I believe affirmative action is problematic in our society because we have demanded that it create parity between the races rather than insure equal opportunity. Preferential treatment does not teach skills, or educate, or instill motivation.

Racial representation is not the same thing as racial development. Representation can be manufactured; development is always hard earned.

Blacks can have no real power without taking responsibility for their own educational and economic development. Whites can have no racial innocence without earning it by eradicating discrimination and helping the disadvantaged to develop. Because we ignored the means, the goals have not been reached and the real work remains to be done (Steele, 1990).

Up to 1990, we failed Humphrey's test.

We should not have been surprised. Those who forget history are doomed to repeat it. There was, and is, a long historical precedent for what we experienced. I refer to the Chinese exams.

The most distinctive feature of ancient Chinese government was the famed examination system. Other countries have until recent times, with few exceptions, been ruled by a hereditary aristocracy, a priesthood, a military hierarchy, or a rich merchant class. But in China, beginning during the first long lived empire in the second century B. C., entry into the bureaucracy that governed the country was limited to those who succeeded in passing a series of very strict governmental examinations, based on a thorough knowledge of the Chinese classics.

The tests were organized and administered in a systematic way and first standardized in the year 669. This was done by the Empress Wu Tse-t'ien, and used as a political weapon. In 690, she became the only woman to rule China, as Emperor Wu Chao.

The system was efficient, and endured for centuries because it favored the selection of the best candidates.

In practice, the examination system operated best in periods of strong political unity, while in times of strife or dynastic change it tended to break down. Likewise, it contained certain manifest defects, such as its undue stress upon memory, and the fact that the wealthy naturally enjoyed superior opportunities to acquire the education that would make success possible.

After the Mongol invasion, in 1210, the exams languished since the Mongols had no proper administrative organization at all. "When the first doctoral examinations were organized in the Chinese style in 1315, quotas were reserved for different nationalities: Of a total of 300 appointments, a quarter were reserved for Mongols, a quarter for foreigners, a quarter for candidates from North China and a quarter for southern Chinese. The operation was a parody of the Chinese competitions, for the Mongols and various foreigners were uneducated and most of the families of literati lived in the lower Yangtze towns, in South China" (Gernet, 1982).

As the sole means of access to political honors and responsibilities, the examinations served to inculcate in those who sat for them the virtues of devotion and submission indispensable to the autocratic empire. They also drained the energies of generations. The artificial character of the tests had been clear since the institution of the composition in eight parts, a sterile, empty stylistic exercise.

The Chinese also had a system for the promotion of officials, including a system of recommendations which made the author of the recommendation jointly responsible for the faults and errors of his protégé.

The successful examinee or promotion candidate was indebted for his success to his parents, to his friends, and sometimes to those who had bet on his success and financed his studies (Gernet, 1982; Bodde, 1981).

The examinations remained in place until 1905, and one can see in this brief account all of the same issues we wrestled with in 1990.

In the United States, our experience with tests dates from 1917, from World War I, with the creation and use of the Army Alpha. Between December, 1917, and November, 1918, 1.7 million men took the Alpha. Letter grades were assigned, ranging from A to D-. Some 8,000 recruits with D- grades were discharged immediately for "mental inferiority". Another 10,000 were assigned to heavy labor battalions. On the high end, those with grades of A, scores of 135 or above, were sent to Officer Candidate School.

The Alpha was used enthusiastically and uncritically. Not surprisingly, the first cases of test misuse occurred with it. Current officers were required to take it. If they did not get a grade of A, they were stripped of their commissions and served out the war as enlisted men.

Let us heed the lessons of history, both Chinese and our own. Most especially, let us listen to those who

speak courageously, publicly, and truthfully. This is one thing we should have done.

Another thing we should have done is to pay more attention to assessing work force 2000.

The 14th annual IPMA Assessment Council Conference had an excellent program, with many important addresses, panels and symposia.

The same month that conference was held, *Personnel Psychology* published a special issue on Project A, the U.S. Army's longitudinal research on personnel selection and placement. Project A was, and still is, the largest, most comprehensive examination of the measurement and meaning of human differences ever undertaken. While many found it's attention to the development of new predictors to be its greatest interest, I found its refinement of criterion measures, and especially the identification of five criterion constructs, to be its finding that moved our field ahead the most. Everyone knows that job performance is multidimensional, except that we continue to treat it unidimensionally. Is it any wonder, then, that members of the Congress treat it unidimensionally, and dichotomously? Project A showed the way for the '90s.

One other aspect of assessment needs comment. As a field, we have neglected and minimized the role of face validity in public acceptance of our work. This has been, and continues to be, short-sighted and unfortunate, and reminds me of the Chinese exams and experience.

Besides paying more attention to assessing work force 2000, we should have paid more attention to training. Indeed, we should have created, not educational reform, but an educational revolution.

We knew there were problems with our educational system in the '60s, '70s and '80s. The problems got lip service. In the '90s they became impossible to ignore.

In 1990, 25 percent of young people dropped out of high school, 70 percent of high school seniors couldn't write a letter applying for a job, and 60 percent couldn't correctly add up their own lunch bills. (Anon., 1990.)

The National Assessment of Educational Progress found that among their 21- 25-year-old sample, only 25% of whites, 7% of hispanics, and 3% of blacks could correctly interpret a complex bus schedule.

But it wasn't only the kids who had problems. The vast majority of American adults could not program their own VCRs.

Adult literacy was a problem, too.



**HOW MANY OF THESE 25 COUNTRIES  
HAVE ADULT LITERACY RATES  
LOWER THAN THE UNITED STATES?**

Argentina	Hungary
Australia	Ireland
Austria	Italy
Belgium	Japan
Britain	New Zealand
Canada	Norway
Chile	Poland
Cuba	Rumania
Czechoslovakia	Spain
East Germany	Sweden
Finland	USSR
France	West Germany
Holland	

*The Economist, 1990*

There were calls for educational reform, including a presidential summit, but the problem was vast and complex. It probably got part of its start with the decisions by many states to fund public education partially from the proceeds of state lotteries. From my perspective, however, "taking a *chance* on education" undermines the moral foundation of the entire venture.

Incidentally, how many of these 25 countries have lower literacy rates than the U.S.? None!

There were signs of hope in 1990, such as in the work done by the National Association of Secondary School Principals, designed to strengthen the leadership skills of principals. A program named *Leader 123* consisted of three phases: *Preparation*, including self- and observer-assessments, *Practice*, during a three day development center (an assessment center with immediate feedback and coaching), and *Performance*, an on-the-job development project with consultation from a job performance coach. Work at the Center for Creative Leadership also drew favorable notice.

Another sign of hope came in a report by the Governor's Commission on Socially Disadvantaged Black Males, appointed by Richard F. Celeste of Ohio. In announcing the commission, Celeste said:

"Simply put, we have wasted too much time. Wasted time for these men, wasted time for our economy. We are now confronted with the inescapable truth--not preached from the busy pulpits of social reform, but printed out in the hard data of computer reports: that social justice has moved beyond the moral imperative. Today it is the economic linchpin of competitive states and nations....

"It is up to us to offer these young men both the challenge and the support they need to contribute their talents as productive and creative members of our society" (Celeste, 1989).



**OHIO'S AFRICAN-AMERICAN MALES:  
A CALL TO ACTION**

Volume One  
Executive Summary



Report of  
The Governor's Commission on  
Socially Disadvantaged Black Males

Of course the professional literature had plenty of ideas, too, concerning training, and also larger issues of organizational psychology.

But education and training wasn't all we should have done. We needed more research. There were severe national problems.

**National Problems**

Education  
Worker Productivity  
Drug Abuse  
Violence  
AIDS  
Physical Health  
Mental Illness  
Aging  
Dependent Care

David Hamburg, president of the Carnegie Endowment, wrote:

In a world full of hatred, repression, terrorism, small wars, and preparation for immense wars, human conflict is a subject that deserves the most careful and searching inquiry. The stakes are now so high that there is an urgent need for cooperative engagement with these problems over a wide range of inquiry involving the physical, biological, behavioral and social sciences.

The time is ripe for the scientific community to provide worldwide leadership in addressing the ubiquity of prejudice, the profound and pervasive impact of ethnocentrism, and the greatly enhanced risks of these ancient orientations in the rapidly changing world of the late 20th century (Hamburg, 1986).

In response to the nation's problems, the leaders of the American Psychological Society convened a summit meeting of some 65 organizations, a meeting which set up a process to create a national research agenda to help focus research efforts and research funding.

## National Research Agenda for Psychology

Education, Training, and Performance  
Health  
Brain, Mind, and Behavior  
Family, Groups, and Community

The research themes for psychology are quite broad, and a lot could fit under any of them. In the education, training and performance theme, I would lobby for research on not just assessment and training, but also on questions of conflict resolution, values, motivation, and service. I wish to especially emphasize service.

In the '80s, the watchword was "If it's not broken, don't fix it." By 1990, you and I had just endured a decade-long frenzy of government bashing, by the leaders of our government, of all people! Public service was regarded as an oxymoron. Bureaucracy came in for its share of knocks. Craig Whitney wrote in the *New York Times*:

"For any politician who cannot deliver on a promise, 'the bureaucracy' is a wonderful scapegoat. So guess who's to blame for Mikhail Gorbachev's failure to dampen a growing economic crisis in the Soviet Union?"

In any nation, the bureaucracy shies away from anything resembling "new thinking". Lenin wrote in 1897 about the complete lack of rights of the people in relation to government officials and the complete absence of control over the privileged bureaucracy.

So after the revolution, did Lenin get rid of it? Of course not. As he observed himself in 1921: "The czar can be sent packing, the landed proprietors sent packing, the capitalists sent packing. That we did. But bureaucratism cannot be 'sent packing', cannot be 'swept from the face of the earth.' One can only reduce it by slow, stubborn effort" (Whitney, 1990).

We Americans, and especially those of us in public service, need to take "service" to heart. Public service is a noble calling.

In the private sector, some service companies are standing their organizations on their heads, making everyone from the chief executive to filing clerks "work for" customer contact employees, helping them to make the most of their brief customer contacts. The best structure for a service firm may end up looking something like a spider's web, with each employee able to tap into the firm's collective knowledge and experience via networked computers. Might such ideas apply in public service?

We also need to learn how better to measure service, and other non-economic dimensions. Economic measures give incomplete and distorted pictures of human welfare. Through the ages, philosophers and religious leaders have denounced materialism as a viable path to human fulfillment. Yet societies across the ideological spectrum have persisted in equating quality of life with increased consumption. Personal self-worth

typically is measured by possessions, just as social progress is judged by GNP growth.

In 1990, the United Nations Development Program defined human development as "a process of enlarging peoples' choices". It created an index, one which goes beyond gross national product, to reflect in a quantitative way the richness of peoples' choices. The index is comprised of gross domestic product per capita, life expectancy at birth, and rate of adult literacy. The human development index provides a measure, one which should have shattered some of our complacency. It is an imperfect index, but it is better than gross national product. If we are going to compete with other nations, we ought to pick worthy pursuits to contest. Incidentally, the United States does not fall in the top 10% of the world's countries on this index.

Turning now to the big picture, another fact that became increasingly clear by 1990 was the growth of our global interdependence. Philip Abelson wrote in *Science*:

Very large volumes of capital, goods, services, components, people, data, and technological know-how flow across national boundaries every day. U.S. businesses participate in a substantial part of the action. They have more than \$1.2 trillion in assets abroad, and in 1988 a third of U.S. multinational companies' earnings came from overseas operations. U.S.-owned operations abroad employ 6.2 million people and depend heavily on local production and technical and management capabilities. The movement of companies is not one-sided. Today more than 3 million Americans work for foreign multinational corporations with affiliates in the United States (Abelson, 1990).

It was clear in 1990 that we in the United States were, as we still are, utterly interdependent with the nations of the world. After World War II, trade, travel and communications increased dramatically. Technology, engineering and science grew explosively. What was little realized is that the human species had become transformed, and interconnected in a way never before known. The change is well expressed in the words of Robert Muller, an assistant secretary-general of the United Nations:

The place of the human being in the grandiose scheme of creation has changed. We ourselves have expanded it. What we have done in reality, and I think this is a thing of the first importance, is that we as a species with limited eyesight, limited hearing, limited brains, limited hands and feet; we have multiplied our eyesight into the infinitely large and the infinitely small, through telescopes and microscopes; we have multiplied the capacity of our hands with machines; we have multiplied the capacity of our feet with airplanes and trains and boats; and we have enlarged the capacity of our hearing with telephones, telecommunications, radio communications, and we are also able to televise things from the other side of the planet; and we have enlarged the capacity of our brain through

incredible computers and machines. So, for all practical purposes, I would beg you to understand that we have become, through science and technology and the search for knowledge, a different species (Muller, 1985).

In short, Muller argues that we are interconnected in a global brain. For the first time, we are part of an organism that can comprehend its own complexity. Surely our comprehension is incomplete, but a fundamental change has occurred during our lifetimes.

It is from this new, broadened, global perspective that we must carry on our work. We needed to build a sustainable world. We needed to promote opportunity, as the way to reduce racism and sexism.

So, that is how it looks in hindsight, from 2001. We needed to speak the truth, to advance our science and our practice of assessing and developing talent, and to become more aware of our interdependence.

=====

Now, please roll back your watches to June 25, 1990.

The good news is that it is not too late to do something about the challenges of the '90s. It's clear that something is broken, and it needs to be fixed.

There is a Chinese proverb which goes: "Seek not to follow in the footsteps of the men of old; seek instead what they sought." It is good advice for us now. We need to embrace our core values.

I have another quote for you from history. It is from the New York Times, published on December 31, 1899: "Tomorrow we enter upon the last year of a century that is marked by greater progress in all that pertains to the material well-being and enlightenment of mankind than all the previous history of the race and the political, social, and moral advancement has been hardly less striking" (New York Times, Dec. 31, 1899).

At the start of the '90s (the 1990s, that is), I would give us a high pass on material well-being, enlightenment, and political advancement. I would give us a pass on social advancement. I would grade us fail on moral advancement.

But it is not too late to save the day. Just as Hubert Humphrey organized the Senate liberals, we need to get organized. We need to preach to the choir, as I have today. But we also need to get out there and spread the word. There's plenty to be done, and we can have some fun doing it along the way. Let's get on with life, liberty, and the pursuit of happiness.

We need to start today. The big decisions of the world are made up of thousands of small ones. Each task is significant, and each choice makes a difference. Keep the core values before you in each choice, and the total will add up, day by day, to a tremendous impact on America and the world.

And by the way, don't wait for empowerment. You're in charge. Just do it!

## References

- Abelson, Philip H. 1990. Science and technology policy. *Science*, 248, 421.
- Anon., 1990. Technical & Skills Training News, Winter, p. 7.
- Bodde, Derk. 1981. *Essays on Chinese civilization*. Princeton, NJ: Princeton University Press.
- Celeste, Governor Richard F., at the signing of Executive Order 89-9, creating the Governor's Commission on Socially Disadvantaged Black Males, April 7, 1989
- Conlon, Charles A. 1890. Recent economic tendencies. *Atlantic Monthly*, 85, 737-748.
- Gernet, Jacques. 1982. *A History of Chinese Civilization*. New York: Cambridge University Press.
- Hamburg, David A. 1986. New risks of prejudice, ethnocentrism, and violence. *Science*, 231, 533.
- Havel, V. 1990. *Disturbing the Peace*. New York: Harcourt Brace Jovanovich.
- Humphrey, H. H. 1976. *Education of a Public Man*. New York: McGraw-Hill.
- Robert Mueller, 1985, Blueprint for a global community. Arlington, Va.: Soundworks.
- Steele, Shelby. 1990. A negative vote on affirmative action. *NY Times Sunday Magazine*, May 13, p. 46 ff.
- Whitney, Craig R. 1990. The revolution is today, bureaucracy is forever. *NY Times*, April 15, p. E 3.



A N I N T R O D U C T I O N T O T H E C L O Z E  
T E C H N I Q U E O F R E A D I N G S K I L L S  
A S S E S S M E N T

Michael J. Dollard  
New York State Department of Civil Service  
Bldg. 1, Harriman Office Campus  
Albany, NY 12239  
(518) 457-2483

The purpose of this paper is to present an introduction to the CLOZE technique of reading skills assessment, how it's done, and its research and theoretical support.

Any discussion of the CLOZE procedure must begin with a consideration of the nature of comprehension. While CLOZE has been used in one form or another since the late 19th Century, contemporary psycholinguistics provides the best theoretical underpinnings for its use.

The basic psycholinguistic construct is that language consists of a surface structure (the physical characteristics of language, i.e., the sound or symbols that exist external to the human being) and a deep structure (the meaning that exists internal to the human being). The mediating forces between the surface structure and the deep structure are syntax, grammar and semantics. The semantic information is gained from experience and exists in a cognitive structure of interrelated facts and relationships. Syntax and grammar are learned rules for the construction and interpretation of language, and are also part of the cognitive structure of the individual.

A second fundamental psycholinguistic construct is that language comprehension is an automatic and continuing activity of human beings, which uses a process of prediction based on the information contained in the cognitive structure of the individual. ("Prediction" in this sense means prior elimination of unlikely alternatives.)

### Dollard - CLOZE Technique\*\*\*

These constructs are laid out reasonably well in Frank Smith's book *Comprehension and Learning* (1975) pp 83-117. Smith is a recognized authority, and is quoted widely in the educational literature. His other books dealing with this area are *Psycholinguistics and Reading* (1973) and *Understanding Reading* (1978).

The CLOZE procedure consists of systematically deleting words from a text, and then having the examinee replace the missing words using clues from the context. Taylor, the modern developer of the CLOZE technique, is quoted in *CLOZE Readability Procedure* (Bormuth, 1967) as indicating that there are three legitimate item construction techniques:

- 1) every n<sup>th</sup> word deletion
- 2) random deletion
- 3) class of word deletion (e.g., delete 20% of the nouns)

Taylor did indicate that the deletion must be a systematic and "objective" method; removing "selected" words renders the technique no more than an ordinary knowledge-based completion test. In its "classical" form, the examinee must produce from within him-/her-self the word to replace the blank (i.e., the "exact word" approach). A more recent development is the MAZE procedure, which is the use of a CLOZE type passage, but with alternative replacement words provided in a multiple choice format.

On the next page (and on the slide) is a page of sample questions for a multiple-choice CLOZE test.

Dollard - CLOZE Technique\*\*\*

SAMPLE PASSAGE DIRECTIONS: The passage below contains 5 numbered blanks. Read the passage once quickly for overall sense. Read it a second time, this time thinking of words that might fit in the blanks. Below the passage are sets of words numbered to match the blanks. Pick the word from each set which seems to make the most sense in that blank as well as the total paragraph. (The size of the blank space is not related to the length of the word that is missing.)

SAMPLE PASSAGE: For determining "late filing", you are to use the following procedure. You should (S1) five working days after the February 28th due date (S2) examining envelopes for late filing. On the sixth working (S3), envelopes must be reviewed for U.S. Postal Service postmark. (S4) from postage meters are not acceptable. Postmarks which are (S5) later than February 28 are considered to be late filed. All envelopes must be retained, showing company name, address, addressee and postmark.

- S1 A. give  
B. make  
C. do  
D. allow

EXPLANATION FOR SAMPLE QUESTION S1:

Choice A--To "give five working days" indicates use for a specific activity. You give time to a project, not to a waiting period.

Choice B--To "make five working days" is not meaningful. You can "make time" for something or to do something but you would not make a time period.

Choice C--"Working days" is not something one can "do".

Choice D--To "allow" is to let happen. In this case, five working days should go by after the due date before examining envelopes for late filing. Choice D is the BEST ANSWER because it has the word which best fits the meaning of the sentence.

- S2 A. since  
B. before  
C. with  
D. by

- S4 A. Parts  
B. Filings  
C. Envelopes  
D. Dates

- S3 A. day  
B. event  
C. stamp  
D. schedule

- S5 A. sent  
B. dated  
C. issued  
D. made

Key: S1 - D, S2 - B, S3 - A, S4 - D, S5 - B

## RESEARCH SUPPORT FOR THE CLOZE TECHNIQUE

Rankin (with Bormuth one of the main writers dealing with readability and CLOZE procedures) produced a seminal paper in "The Validity of CLOZE Test in Relation to a Psycholinguistic Conceptualization of Reading Comprehension" in which he said, "Reading Comprehension, viewed from a psycholinguistics perspective, is a process of mapping new information into existing knowledge structures and is similar to general thinking, reasoning and problem solving. Reading comprehension is an active process -- a dialogue between the reader and writer -- and an integration of the printed page with the reader's mind ... Further research and a review of the literature indicate that the content validity of CLOZE tests as measures of a meaningful construct of comprehension is stronger than for conventional reading comprehension tests ... Furthermore, the construct validity of reading comprehension measured by CLOZE tests is more theoretical and empirically meaningful than notions of reading comprehension embodied in conventional reading tests. ... The large amount of research on CLOZE test criterion-related validity demonstrates high positive correlations between CLOZE test results and other reading test data indicating that CLOZE tests and conventional tests are measuring something in common ..."

Rankin goes on to make some assumptions about reading comprehension which are consistent with the psycholinguistic definition (as in Smith, above, and in Pearson and Johnson, 1978 and Oller, 1979) together with observations on the relationship of CLOZE to these assumptions:

# Dollard - CLOZE Technique\*\*\*

"1. Reading comprehension as a process involves the extraction of information from a written source prior to the utilization of this information. . . .

"CLOZE tests intermittently sample the inferences (syntactic and semantic) the reader is making during the process of reading the passage. Thus CLOZE scores become an indicator of the reader's ability to process information in the text, not the ability to respond to specific questions which may not have been of any concern to them at the time when they were reading the text. [ed. The problems referred to here include that of asking hard questions about easy text, and asking easy questions about hard text.]

"2. Reading comprehension is an active process -- a dialogue between a reader and a writer -- in which the effectiveness of communication is dependent upon the expectancies of both the writer and the reader. . . .

"CLOZE tests repeatedly sample the likenesses of the pragmatic expectations of the reader and the writer and, thus, sample the effectiveness of communication."

"3. Drawing inferences is essential to the processing of information through reading. . . .

"This redundancy in print is categorized as sequential or distributional. Sequential redundancy includes the syntactic and semantic constraints imposed by the language. That is, certain types of words follow other



### Dollard - CLOZE Technique\*\*\*

words, hold certain positions, and fulfill certain functions within a sentence. Distributional redundancy refers to the fact that some words occur more frequently in the language, regardless of context, than do other words ... Skilled readers have both sequential and distributional redundancy rules programmed in their long term memory." . . .

Rankin finishes his observations by saying, "Although CLOZE items might look like 'fill-in-the-blank' items calling only upon rote memory, they most often measure inferential processes. It seems clear [from the examples Rankin offers in the paper] ... that CLOZE is sensitive to the higher levels of inferences as categorized by Pearson and Johnson (1978)

David Wardell in his Portland State University dissertation *CLOZE procedures: An Historical Review* similarly states "Redundancy is placed in a verbal context and can be noted on three separate levels of language:

- 1) surface syntactic structure
- 2) deep syntactic structure
- 3) semantic structure.

Each level is said to contribute to an individual's ability to perform with CLOZE materials." He then goes on to note that Bachman (1982) suggests that a modified CLOZE passage, using rational deletions [i.e., deleting specific words on a subjective basis, in opposition to the position of Taylor, above] is capable of measuring both syntactic and discourse level relationships within a text, and that Ellington (1981) specifies a number of areas that can be addressed by the CLOZE procedures using either random or systematic deletion:

## Dollard - CLOZE Technique\*\*\*

- 1) general context
- 2) content
- 3) process strategies
- 4) specific phonic or morphic elements
- 5) relationships cued by function words, pronouns and pronoun referents
- 6) organizational patterns.

John Bormuth in "New Data on Reliability", an unpublished paper presented to AERA in Seattle in 1967, summarized the research to that date, and while additional aspects of language have since been added to his list as shown below, the summary still seems apt:

"The CLOZE readability procedure immediately drew the attention of readability researchers who set about studying CLOZE tests to see if they were valid and reliable measures of the comprehension difficulty of passages. This research has become too extensive to review here. Bormuth (1967) and Rankin (1964) have each published detailed analyses of this research. In general, research showed that CLOZE readability tests are highly valid and highly reliable measures of the comprehension abilities of students and of the comprehension difficulties of materials." He then went on to list some of the aspects of language which research had shown can be appropriately evaluated by CLOZE tests [NOTE that this list was compiled prior to most of the major work in psycholinguistics, and represents a much more primitive categorical structure. mjd]:

vocabulary complexity

word length

### Dollard - CLOZE Technique\*\*\*

- morphological complexity
- abstractness
- frequency
- grammatical complexity
- syntactic depth
- modifier distance
- transformational complexity.

Wardell reports Bormuth (1969) as studying "the factor validity of CLOZE tests as measures of comprehension ability by analyzing the principal components of the correlations among nine CLOZE tests and seven multiple choice tests, each designed to measure a different comprehension skill. His results showed that CLOZE tests seem to offer a valid, convenient and completely objective method of constructing tests that can be used for measuring either the reading comprehension abilities of students or the comprehension difficulty of passages. Rankin reports a factor analytical study by Kohler in 1966 that found logical reasoning, wide-range vocabulary and inference to be consistently and significantly related to CLOZE tests.

Bormuth in a 1967 UCLA Research Report (CSE IP - OR - 1) reports data by Taylor (1953) showing that CLOZE readability test difficulties ranked the passages in the same order the readability formulas ranked them. In a 1964 UCLA Research Report Bormuth reports "Taylor (1953) found a significant correlation between the CLOZE test difficulties of a set of passages and the readabilities of the passages as predicted from the Dale-Chall and Flesch formulas. Further, he found that the CLOZE tests

### Dollard - CLOZE Technique\*\*\*

measured elements of style which affect difficulty, but to which the formulas are insensitive." He also reports there a study by Sukiori (1957) which found a correlation of .83 between the judged difficulties a set of passages and their difficulties as measured by CLOZE tests.

Sticht, in his book *Reading for Working: A Functional Literacy Anthology*, says "Research has indicated that, although there is no single definitive method for measuring reading comprehension, the 'mechanical' CLOZE procedure has consistently yielded very high correlations with multiple-choice tests and other more subjectively constructed measures of comprehension and difficulty. Therefore the weight of the evidence indicates that the cloze test provides a *valid* measure of reading comprehension. The fact that it is also strictly objective, and that  $n$  independent alternative forms can be created simply by deleting every  $n^{\text{th}}$  word counting from the first, second, ..., or  $n^{\text{th}}$  word from the beginning of the passage, further encouraged the use of the cloze procedure." Sticht bases this conclusion on his understanding of Taylor (1953), Bormuth (1969) and Rankin and Culhane (1969).

The research literature seems compelling that CLOZE tests are capable of evaluating reading comprehension ability with at least as much precision and validity as conventional multiple-choice tests, and that they evaluate syntactic, grammatical and semantic features of language. Further, studies of the Degrees of Reading Power test using Rasch and other methodologies show it to be culture fair. Intergroup differences on CLOZE tests studied in England [CLOZE Reading Tests, published by Young, Hodder & Stoughton, and reported in the Ninth Mental Measurements Yearbook] also

### Dollard - CLOZE Technique\*\*\*

showed reduced inter-group differences. Bormuth in his 1967 UCLA report concludes that "The CLOZE readability procedure has a number of advantages not shared by other available methods of determining [comprehension] difficulty. Unlike the conventional test item used in other methods whose materials are tried out directly on students, CLOZE test items are easily made and do not inject irrelevant sources of variance into the measurement of difficulty."

### CONSTRUCTING CLOZE TESTS

Accepting from the above that CLOZE tests are a valid means of evaluating reading comprehension, the question then becomes what is the appropriate methodology for developing the tests. This question resolves to two fundamentally different sub-questions:

- 1) how do you select the texts that will form the basis of the tests?, and
- 2) how do you select the words to delete from the text?

Unfortunately, the literature does not address the first sub-question. The research referenced above clearly indicates that CLOZE procedures will evaluate whatever language features are present in a text, and contain some feature lists that have been studied. Dollard in his *Reading Skills Assessment: A Survey of Practice* (1989) has used a feature list covering syntactic comprehension, literal comprehension and interpretive/evaluative comprehension. Survey response from public sector employers identifies most of these features as being evaluated by the conventional multiple-choice tests being used in employee selection and -- presumably --

### Dollard - CLOZE Technique\*\*\*

existing in the texts used at public sector work sites (virtually all respondents report basing their conventional multiple-choice test items materials drawn from the work site). The work of Francis and Kucera (1982) on the Brown University Corpus would suggest that the distribution of these language features will vary by text type and context. The implication of this seems to be that in using CLOZE technique in an employment selection context, care has to be taken to insure that the texts used for evaluation purposes sample on a representative basis the language features found in the texts used on the job.

The literature on the methodology to be used in actually creating the CLOZE test is fortunately more illuminating. Bormuth (1967) reports Taylor -- the modern developer of CLOZE tests -- as saying in 1953, "While nearly all readability research employs tests made by deleting every 5<sup>th</sup> word, CLOZE tests can be made by deleting every n<sup>th</sup> word, words at random or just the words of a given type. The only restriction is that the words deleted must be selected entirely by an objectively specified process, otherwise the test must be classified as a common completion test." In the same monograph, Bormuth quotes Taylor in a 1955 source as qualifying this statement to the effect that deletions of the third type (i.e., words of a given type) are rarely practical because of the frequency distribution of most word classes in text. Despite Taylor's prohibition against deleting subjectively chosen words on a non-systematic basis, Wardell does report a study by Bachman (1982) which suggests that such "rational deletions" are capable of measuring both syntactic and discourse level relationships. However, research by McGinitie (1961) suggests that every n<sup>th</sup> word deletions work best.



### Dollard - CLOZE Technique\*\*\*

In a study reported to the 1982 Annual Meeting of the World Congress on Reading, Templer provides rules for constructing CLOZE tests:

- "1) Select two passages. Allow for deletion of 25 words in each passage with one sentence unmutilated at the beginning and end of each passage.
- 2) Delete words in a regular pattern (no less than every 5<sup>th</sup> and no more than every 10<sup>th</sup> word -- research indicates that the word deletion rate within these limits does not affect results or reliability)
- 3) The proportion of nouns deleted must be approximately equal to 20% of all nouns in the passage -- Do not delete proper nouns or numerals unless they can be deduced from the context. Contracted or hyphenated words count as one word. Retain all punctuation."

Templer comments further on the proportion of nouns deleted: "This is very important because it affects the validity of the test. It also makes a difference to the difficulty level of a passage if mostly prepositions, pronouns, verbs and conjunctions are to be replaced rather than nouns, because nouns are more difficult to replace.

The number of items/deletions is important as it affects reliability. Templer, above, recommends two passages, each with 25 deletions. However Templer's task was the scaling of textbooks, in an educational setting and the matching of student comprehension ability to text comprehension difficulty.

CONCLUSION

### Dollard - CLOZE Technique\*\*\*

The CLOZE technique appears to be a cost-effective and valid way to assess reading skills. It has strong theoretical support in psycholinguistic theory, and very strong empirical research support in the education field. It is far from rote "fill-in-the-blank" knowledge recall; it does seem to validly assess the full range of reading skills.

\*\*\* BIBLIOGRAPHY \*\*\*

- , *DRP Handbook*, The College Board, New York, NY (1988)
- , *Ninth Mental Measurements Yearbook*, Buros Institute, University of Nebraska (Reading Tests)
- Bormuth, J. *Relationships Between Selected Language Variables and Comprehension Ability and Difficulty* UCLA Research Report #2082 (1964)
- Bormuth, J. *Cloze Readability Procedure* UCLA Research Report (CSE IP-OR-1) (1967)
- Bormuth, J. "New Data on Reliability" unpublished paper presented to AERA, Seattle, (1967-2)
- Bormuth, J. *Development of Readability Analyses* University of Chicago, Final Report Project 7-0052 HEW Contract OEC-3-7-070052-0326, March 1969
- Dollard, N. *Reading Skills Assessment: A Survey of Practice* New York State Department of Civil Service (1989)
- Koslin, B.; Zeno, S. and Koslin, S. *The DRP: An Effectiveness Measure in Reading.*, The College Board (1987)
- Rankin, E. and Culhane, J. "Comparable Cloze and Multiple-Choice Comprehension Test Scores" *Journal of Reading* No 13 pp193-198 (1969)
- Rankin, E. "The Validity of Cloze Tests in Relation to a Psycholinguistic Conceptualization of Reading Comprehension", *Forum for Reading*"
- Smith, F. *Comprehension and Learning* (1975)
- Smith, F. *Psycholinguistics and Reading* (1973)
- Smith, F. *Understanding Reading* (1978)
- Sticht, T. ed. *Reading for Working: A Functional Literacy Anthology*
- Taylor, W. "Cloze procedure: A New Tool for Measuring Readability" *Journalism Quarterly* vol 30 pp 415-433 (1953)
- Templer, Lois. *Readability: Cloze Procedures for Assessing High School Text and Resource Books.* unpublished report to the Annual Meeting of the World Congress on Reading, Dublin (1982)
- Wardell, D. *Cloze Procedures: An Historical Review* Portland State University Dissertation

---

## **Employee Referral Programs: Do Successful Employees Refer Better Applicants Than Unsuccessful Employees?**

**Michael G. Aamodt and Glen T. Rupert**  
**Radford University**

---

Personnel professionals have long been interested in the best ways in which to recruit potential employees. This interest stems from two main ideas. The first idea is that certain recruitment methods will yield higher numbers of acceptable applicants, thus making the recruitment process less expensive (Kirnan, Farley, & Geisinger, 1989). For example, if a \$100.00 newspaper advertisement results in 50 applicants for a job compared to two applicants resulting from a \$3,000 fee paid to an employment agency, then an organization might be better off recruiting through newspaper ads.

The second idea is that certain recruitment methods will attract employees who, once on the job, perform better than employees recruited by other methods. That is, even though newspaper ads in the previous example yielded more applicants, it is possible that none of the 50 will perform as well or stay with the organization as long as would the two from the employment agency. Thus, the savings obtained in recruitment costs would be nullified by the increased training expenses and the reduction in employee performance.

Though several studies have reported that certain recruitment methods yield better applicants than do others (e.g. Breaugh, 1981; Decker & Cornelius, 1979; Reid, 1972), the meta-analysis by Aamodt and Carr (1988) reveals that employees hired as the result of an employee referral have longer tenure than do employees recruited through other means. However, the meta-analysis also indicates that recruitment methods appear to be equal in regard to employee performance.

Even though employee referrals are superior only when tenure is used as the criterion, several theories have been postulated about why referrals result in better employees. The first of these theories suggests that applicants who are referred by other employees receive more accurate information about the job than do employees recruited by other methods (Wanous, 1980). In essence, the applicant receives a realistic job preview from a current employee. This theory has not only received some support in the literature (Breaugh & Mann, 1984; Quaglieri, 1982) but also is consistent with the results of the Aamodt and Carr (1988) meta-analysis. That is, employee referrals were only superior in regard to tenure and realistic job previews have also been found to have their greatest effect on tenure rather than performance (Premack & Wanous, 1985).

The second theory postulates that differences in recruitment source effectiveness are due to the fact that formal and informal sources reach and are used by different types of applicants (Schwab, 1982). While some research has supported this theory (Breaugh & Mann, 1984; Ellis & Taylor, 1983; Taylor & Schmidt, 1983; Swaroff, Barclay, & Bass, 1985), other research has not (Aamodt & Carr, 1988; Breaugh, 1981). Furthermore, Aamodt and Carr (1988) pointed out that results are not consistent across studies and that applicants tend to use a variety of recruitment strategies rather than just one.

Though the two theories mentioned above have received at least some empirical support, there is a third possibility which might better explain the finding that employee referrals result in greater tenure than do the other recruitment strategies. This third theory (Aamodt & Carr, 1988) has its roots in the interpersonal attraction literature which indicates that people tend to be friends with others who are similar to themselves (Byrne, 1971). If this is true, and the research strongly suggests that it is, then an employee recommending a friend for a job will more than likely recommend a friend similar to him/herself. Thus, it would make sense that a person who is happy with their job would recommend a person who, due to his/her similarity to the incumbent, should also be happy with the job. Likewise, it would make sense that an unhappy employee would recommend similar friends who would also be unhappy and would probably have a short tenure with the company.

If this theory is true, then it is not the employee referral process per se that makes the difference in the future success of an applicant. Instead, the success of a future employee is a function of the person who makes the referral. It is the purpose of this study to test this idea using four separate samples of employees who were recruited through employee referrals by comparing the tenure and performance of employees who referred an applicant with the subsequent tenure and performance of the applicants they referred.

## Method

### Subjects

The first sample consisted of 135 former retail employees who had been employed part-time and who had been referred for the job by a friend. The mean tenure for the first sample was 9.8 months with a standard deviation of 11.75 months.

The second sample consisted of 29 male employees who had been referred by a friend and worked full-time for a company that installed fire protection equipment such as automatic sprinkler systems. The mean tenure for the second sample was 12.39 months with a standard deviation of 18.09 months.

The third sample consisted of 24 male employees who worked for a concrete manufacturing company. Each employee had been referred for his job by a fellow employee.

The fourth sample consisted of 42 former restaurant workers who had been referred for the job by a friend. The mean tenure was 9.2 months with a standard deviation of 11.5 months.

### Procedure

For the first and fourth samples, tenure information was obtained for both the employee who made the referral as well as the employee who was referred. The tenure of the employee making the referral was then correlated with the eventual tenure of the employee who was referred. Unfortunately, no performance data were available for these samples.

The same procedure was used for the second sample that was used for the first sample. However, for the second sample, performance ratings were made on a

performance appraisal form already used by the company. This form, completed by the employee's supervisor, contained 10 items and most resembled a behavioral rating scale. The maximum score that could be earned by an employee was 200 with a mean score of 153.78 and a standard deviation of 63.02.

The performance ratings for the third sample were also taken from the performance appraisal system currently used by the concrete manufacturer. This appraisal system involved a supervisor evaluating each employee on six categories: Initiative, work quality, work quantity, cooperation, knowledge, and dependability. The maximum total score possible was 18 and the mean for our sample was 14.17 with a standard deviation of 2.33. Tenure data were not available for this sample.

### Results and Discussion

As shown in the table below, the results indicated that for the retail, restaurant and fire protection samples, a significant relationship existed between the tenure of the employee making the referral and the employee being referred. Thus, in terms of tenure, successful employees tend to refer applicants who will also be successful. However, there was no significant relationship involving the performance ratings for either the fire protection agency employees nor the concrete company employees.

Sample	Criterion	
	Tenure	Performance
Retail Workers	.24	
Fire Protection	.70	-.01
Concrete Workers		.33
Restaurant Workers	.25	

These results have obvious implications for organizations using employee referral programs. Rather than treating all referrals as being equally valuable, referrals from long tenure employees should be considered first.

An interesting outcome of this study was the lack of a significant relationship involving performance ratings. While these results are consistent with the cumulative results of previous research indicating that recruitment source is not related to subsequent employee performance, they are not consistent with the idea that successful employees will refer successful future employees.

Perhaps the reason for this lack of a relationship involving performance is that we choose friends who are similar to ourselves in terms of interests, attitudes and personality (characteristics commonly found related to job tenure) but perhaps not in abilities (characteristics commonly found to be related to job performance). Support for this idea comes from Streicher (1989) who found a correlation of .70 between the personality styles of friends and from Smith and Redden (1990) who were unable to find significant correlations between the abilities of friends.

## References

- Aamodt, M.G., & Carr, K. (1988). Relationship between recruitment source and employee behavior. Paper presented at the annual meeting of the International Personnel Management Association - Assessment Council, Las Vegas, Nevada.
- Breaugh, J.A. (1981). Relationships between recruiting sources and employee performance, absenteeism, and work attitudes. Academy of Management Journal, 24, 142-147.
- Breaugh, J.A., & Mann, R.B. (1984). Recruiting source effects: A test of two alternative explanations. Journal of Occupational Psychology, 57, 261-267.
- Birne, D. (1971). The attraction paradigm. New York: Academic Press.
- Decker, P.J., & Cornelius, E.T. (1979). A note on recruiting sources and job survival rates. Journal of Applied Psychology, 64, 463-464.
- Ellis, R.A., & Taylor, S.M. (1983). Role of self-esteem within the job search process. Journal of Applied Psychology, 68, 632-640.
- Kirnan, J. P., Farley, J. A., & Geisinger, K. F. (1989). The relationship between recruiting source, applicant quality, and hire performance: An analysis by sex, ethnicity, and age. Personnel Psychology, 42, 293-308.
- Premack, S.L., & Wanous, J.P. (1985). A meta-analysis of realistic job preview experiments. Journal of Applied Psychology, 70, 706-719.
- Quaglieri, P.L. (1982). A note on variations in recruiting information obtained through different sources. Journal of Occupational Psychology, 55, 53-55.
- Reid, G.L. (1972). Job search and the effectiveness of job-finding methods. Industrial and Labor Relations Review, 25, 479-495.
- Schwab, D.P. (1982). Organization recruiting and the decision to participate. In Rowland, K., and Ferris, G. (Eds.), Personnel management: New perspectives. Boston: Allyn & Bacon.
- Smith, M., & Redden, T. (1990). Relationship between the abilities of friends. Proceedings of the 11th Annual Graduate Conference in Organizational Behavior and I/O Psychology.
- Swaroff, P.G., Barclay, L.A., & Bass, A.R. (1985). Recruiting sources: Another look. Journal of Applied Psychology, 70, 720-728.
- Taylor, M.S., & Schmidt, D.W. (1983). A process-oriented investigation of recruitment source effectiveness. Personnel Psychology, 36, 343-354.
- Wanous, J.P. (1980). Organizational entry: Recruitment, selection, and socialization of new comers. Reading, MA: Addison-Wesley Publishing Company.



Job Analysis: Applications and Innovations

Susana R. Lozada-Larsen, Ph.D.  
The Psychological Corporation

Paper presented at the 1990 IPMA Assessment Council Annual Conference  
on Personnel Assessment, San Diego, California

Virtually every activity of personnel administration -- from the classification of jobs, the selection of employees, the training and development of personnel, the development of performance appraisal systems to the setting of compensation rates -- begins with job analysis. One type of job analysis technique attempts to collect data that can serve as the foundation for all of these major human resource functions. Known as common-metric job analysis, generic job analysis, or structured job analysis, this technique is particularly suited to the functions of job classification, selection, and compensation. Having important implications for the creation of an integrated human resource system, this presentation will present an overview of common-metric job analysis. Two new and innovative common-metric job analysis techniques will be described, with attention paid to how these common-metric job analysis techniques may be used to simultaneously contribute to human resource applications in general, and to job classification, compensation and selection validation in particular.

First and foremost, common-metric job analysis techniques are used to empirically compare and contrast jobs of different types. In order to effectively compare different jobs, common-metric techniques rely upon one set of descriptors; that is one "metric," that is supposed to be applicable, or "common," to all jobs. Hence the term "common-metric" refers to the use of one set of descriptors that will be used to analyze all different types of jobs. Using as little technological specificity as possible, the language of the common-metric approach is worker-oriented, focusing on the domain of observable, general worker behaviors. In contrast to a task-orientation, in which highly specific job activities are described in detailed behavioral terms, the worker-oriented approach describes more general work behaviors. Hence, jobs that differ at the task level can be meaningfully compared using the worker-oriented language of the common-metric approach.

Many human resource functions require the ability to objectively compare and contrast jobs. Setting up a compensation system entails comparing and contrasting jobs in order to produce a hierarchy that reflects the comparative value of jobs on the market or, alternatively, their "inherent" value to the organization. Take for example, the jobs of Fire Fighter and Police Patrol Officer. A task-oriented job analysis would conclude that these jobs have virtually no similarity; that is, one involves law enforcement whereas the other deals with fire fighting. In contrast, a common-metric, worker-oriented analysis would conclude that these jobs are quite similar in terms of having high levels of hazardous job duties, unpleasant work environment, dealing with the public, and vehicle operation, and low levels of supervising employees, communicating decisions, businesslike work situations, and operating precision machines. For compensation purposes, a common-metric job analysis technique would lead to the similar placement of these two jobs in the wage and salary structure.

Another human resource function that depends upon the ability to compare jobs is job classification. In job classification, the primary purpose is to examine the similarities and differences between jobs with the intent of producing coherent groupings of similar jobs. By collecting quantitative ratings of jobs using the same set of job

descriptors, common-metric job analysis data is ideally suited for empirical job classification analyses. In general, the more jobs an organization has, and the wider the variety of jobs, the greater the need to use empirical job classification techniques as opposed to reliance on expert judgment or title-to-title comparisons. Job classification is seldom undertaken for its own sake. Instead, the way jobs are grouped, or classified, will in turn affect compensation practices, reorganization plans, selection validity investigations, performance appraisal systems, and training and development programs.

Now that we have defined common-metric job analysis and explained in general terms for what human resource applications this technique can be used, let's take a look at an actual common-metric questionnaire. The first common-metric questionnaire is unambiguously titled: The Common-Metric Questionnaire (CMQ). Authored by Dr. Robert J. Harvey and published by The Psychological Corporation, the CMQ is designed to be able to describe, analyze, and compare jobs of all different types -- from heavy equipment operators to office support staff to senior management. The CMQ is designed to correct some of the major stumbling blocks that have plagued common-metric techniques to date. Namely, the CMQ is designed to capture managerial work as well as the work of other occupational families. In addition, written at no higher than an eighth grade reading level, the CMQ has been specially designed for self-administration by job incumbents. This feature, which is unique to the CMQ, frees organizations from having to employ or train professional job analysts, and permits employee involvement in job analysis. Depending upon the literacy level of the employees, self-administration times range from 40 minutes to two hours.

Building on decades of research on job analysis and job evaluation, the CMQ is based on the idea that there are four dimensions, or factors, underlying human work. Taken together, this vitally important set of factors presents a practical and meaningful way to characterize the world of work (see Figure 1). Each of the four factors is represented by a unique set of worker-oriented, behaviorally-based questions, for a total pool of 238 items. Using a matrix type format, each item is further supplemented by a series of special rating scales, resulting in possible grand total of approximately 1670 ratings. Hence, the CMQ can describe, compare, and order jobs using a total range of 238 to over 1600 descriptors. (Overhead example to demonstrate matrix rating format and explain rating task).

If used in the context of compensation, the four CMQ factors serve as the compensable factors on which the compensation structure is based. Job incumbents complete the CMQ, indicating which activities they perform and the relative time spent on each activity. After supervisors or personnel specialists check the ratings for accuracy, a market capturing scoring formula is applied to each rating. In effect, this formula gives the market value of each work activity described in the CMQ. Once an incumbent has indicated which activities are part of their job, the market capturing scoring formula can be used to tally up the current market value of the job. Or, perhaps an organization would rather learn more about their own current valuation policies. By regressing incumbent job analysis data

against current salary, organizations can learn about their internal valuation policies. This type of policy capturing technique allows organizations to pinpoint exactly where their wage and salary dollars are going and learn precisely what work activities they currently value the most.

Another important application of the CMQ is in the area of selection validation. The CMQ is designed to alleviate both the expense and complexity of selection validation. Using the CMQ, organizations can collect all major types of validity evidence, support alternative validation strategies to broaden and expand their selection program capabilities, as well as establish the job-relatedness of their selection procedures.

Organizations invest in selection programs for two primary reasons: (a) to staff the organization with competitive and productive employees in an effort to maintain or gain a competitive edge over rival organizations, and (b) to avoid unlawful discriminatory practices in the choice and arrangement of human resources. To protect the organization's investment in selection, selection programs should be monitored and refined to ensure that they are working to move the most productive employees into and up through the organizational structure.

In order to gauge whether a selection program is operating effectively and legally, there are two primary lines of evidence that may be gathered: validity evidence and evidence of job-relatedness. Validity evidence empirically and logically supports the rationale that people who do well on selection tests will be productive employees. Validity is a reflection of the soundness and appropriateness of this rationale, rather than a reflection of the quality of selection predictors in and of themselves. Job-relatedness evidence serve to confirm the logic and appropriateness of selection procedures. Without validity and job-relatedness evidence, organizations can not be sure that their selection program is actually working. In effect, organizations stand to waste their business investment in selection programs and risk discriminatory practices in the hiring and promotion of human resources.

It is both amazing and disturbing to find so many organizations who are interested enough in the idea of selection to invest in setting up a selection program, but are not interested enough to take the extra step of finding out whether their programs really work. More often than not, validation studies are seen as a token means to satisfy legal requirements and, as such, are conducted with the minimalist of effort and quality. I think this may due in part to a misunderstanding of what validation actually contributes to a selection program, in part due to an underestimation of the business information that can be gathered by a good quality validity information, and in part due to the extra expense and expertise that good validation entails.

Before going on to describe how the CMQ can contribute to each type of validity evidence, I think it is important to point out that validity evidence is not the only kind of evidence that is important to support



a selection program. In general, organizations should be concerned about the quality of all aspects of their selection programs, including the methods used to choose or create predictors, the quality of predictors, and the quality of their job performance, or "criterion" measures. In particular, evidence of the job-relatedness of selection procedures is an essential component in establishing the soundness, appropriateness, and legality of selection programs. In fact, job-relatedness may be the most important ingredient to ensure that selection programs are acceptable to those whom they affect the most: the applicants and the employees.

The issue of job-relatedness is closely aligned with the issue of validation. In fact, it is not uncommon for these two terms to be used interchangeable. However, these are not synonyms. In contrast to validity evidence, job-relatedness is concerned with establishing a logical justification for selection predictors. As such, experts have pointed out that a predictor may be valid, but not job-related, or job-related in instances where the question of validity is not appropriate. One reason for the confusion about the concepts of validity and job-relatedness is that validity evidence is generally seen as a means to establish evidence of the job-relatedness of predictors. To the extent that validity evidence contributes to a logical support of the use of predictors, then validity evidence does indeed establish an argument of job-relatedness. In general, more than one line of validity evidence is necessary to contribute to our logical understanding of the relationship between predictor scores and future job performance. However, if one can substantiate the requirements of job-relatedness by logical argument, further validity evidence may not be necessarily required to establish job-relatedness.

Beginning now with validity evidence: There are three major types of validity evidence. Each one contributes a different type of information about the soundness of the rationale that people who score highly on the selection tests, or "predictors," will be high performing employees. For the most part, it is a good idea to gather as many of these types of validity evidence as possible to support the use of selection predictors. Most organizations don't do this, perhaps because of the expense, or because some of these strains of evidence entail rather complicated procedures and require a fair amount of expertise on the part of in-house selection specialists. In any case, and particularly if a selection program assessment rests on just one type of validity evidence, it is critically important that the validity evidence be gathered according to relevant professional guidelines and standards and be of the highest quality possible.

The first type of validity evidence is criterion-related validity. This is the empirical -- or statistical -- approach to gathering validity evidence. By collecting actual employee job performance scores, or "criterion" scores, and test, or predictor, scores, one can statistically determine how much confidence may be placed in using test scores to predict future job performance. In order for this approach to work properly, and in order to serve as evidence of job-relatedness, it is important that job performance scores be collected from knowledgeable, trained raters using consistent and sound measures of job performance. It is on the criterion -- the job performance --

side of the equation that the CMQ comes in. The CMQ data can be organized to create relevant, reliable, and uncontaminated job performance measures. First, the CMQ is used to either classify jobs into common occupational families or check on the accuracy of current classification structures. For each occupational family, the CMQ data is organized in the form of job family descriptions. Finally, a performance rating metric is merged with job family descriptions to produce job performance measures. CMQ criterion measures delineate important work behaviors, satisfying the critical goal of criterion relevance. Criterion reliability and freedom for contamination are assured by reliance upon actual, observable work behaviors rather than vague abstractions about worker characteristics. The use of supervisor-incumbent dyads as a source of job performance ratings and rater training is recommended to ensure accurate and reliable ratings.

Content validation relies upon logical -- as opposed to statistical -- evidence. Evidence of content validity is gathered by comparing the content of the predictor with the content of the job. If the predictor is a test, then the test questions are compared with work activities. In order to establish evidence of content validity, the test questions should mirror, or "sample," important work activities. In order to make the comparison between predictor content and job content, it is necessary to have a complete description of observable work behaviors. For job specific predictors, a demonstration of content validity can serve as evidence of job-relatedness. By covering over 1600 different kinds of work activities, the CMQ job and job family descriptions present detailed listings of observable work activities. These descriptions allow organizations to effectively make the job-predictor comparison that serves as the focus of content validity investigations.

Construct validation relies upon both criterion-related and content validation evidence, in addition to information about predictor development procedures and the overlap of the predictor with other predictors. Evidence of construct validity demonstrates two things: (a) that the construct measured by the predictor (e.g., verbal reasoning or numerical ability) is required for successful job performance, and (b) that the predictor is a "good" measure of the construct. By contributing to criterion-related validation and content validation, the CMQ helps to give selection researcher information about construct validity. In addition, the CMQ is amenable to synthetic validation, an alternative validation approach that also contributes to construct validity evidence.

Alternative validation strategies are essentially short-cut ways to establish the validity of inferences about predictor scores without having to gather the more traditional validity evidence. Synthetic validity evidence is collected by first analyzing jobs into their component parts. Then, for each component part, traditional validity evidence is gathered for predictors. The primary point of synthetic validation is to map the relationship between job component scores and predictor scores. Once this is done, it is possible to support the use of predictors for jobs without having to carry out empirical validity studies in each new applied setting. Job analysis to break jobs down into their component parts is the only requisite activity



that must be carried out in each new setting in order to use synthetic validity evidence. CMQ job analysis does exactly this: break jobs down into their component parts. By using a synthetic validity approach, CMQ job component scores can be empirically linked with predictor scores. Such empirical linking establishes the relationship between work activities, worker performance, and the possession of relevant KSAO's. Specifically, empirical linking between jobs and predictors can support, or even supplant, subject matter expert judgments of KSAO requirements, support the selection or arrangement of predictors, as well as help to establish job relatedness evidence to support the use of common-selection procedures across task-dissimilar jobs.

In the same vein, the CMQ is also capable of supporting arguments of validity transportability. Transportability is based on the idea that validity evidence gathered for a predictor in one selection program can be used -- or "transported" -- as validity evidence for the same predictor in other selection programs, but only under one condition. In order to transport validity, the jobs for which the validity evidence was originally gathered must be demonstrably the same as the jobs in the current selection program. In effect, it is necessary to conduct job analysis to establish the similarity of these two job groups. First, traditional validity evidence is gathered for a particular set of jobs that have been described using the CMQ. This validity evidence may be used to support the same predictors for a new set of jobs by: (a) analyzing the new set of jobs using the CMQ, and (b) based on dimension-to-dimension empirical comparisons, demonstrating that the new set of jobs is similar to the set of job for which the traditional validity evidence was gathered.

The second common-metric questionnaire I'd like to show you is an occupational-specific questionnaire. That is, this questionnaire employs a set of descriptors that will be applicable to different jobs all belonging to the same occupational family: the managerial job family. More job-specific than the CMQ, this questionnaire can be used in the same way as the CMQ for managerial compensation, classification, and selection, as well as training, development and performance appraisal. The initial field test moniker of this instrument was EXCEL - The Executive Checklist. After some preliminary field testing, EXCEL is currently undergoing expansion work.

The creation of EXCEL began with a research effort to identify and define the common dimensions that underlie managerial work. Based on a combination of interview and observation-based research of public sector managers, management work theory research, and job analysis research, an integrated picture of managerial work was derived. This integrated picture presents and carefully defines seven dimensions common to the work of managers, supervisors, and executives (see Figure 2).

After identifying and defining the seven common dimensions, EXCEL was developed by generating over 1700 items and ultimately field testing 249 worker-oriented, behaviorally-based job analysis items (overhead example). Using EXCEL, managerial job incumbents indicate which of

these items are part of their job. For each item that is part of the job, the incumbent indicates the frequency of the activity, the criticality, and whether the item is delegated, shared with others, or performed personally and singly. Administered to 498 managers, supervisors, and executives of two companies in the Northeastern region of the United States, the results of our initial field testing were quite positive. Solid support for the seven dimensional structure of the instrument was obtained, and our preliminary item analyses indicated omission or revisions of only 14 of the original 249 items.

Since this initial field testing, we have articulated five major design goals for EXCEL to guide the process of expanding the questionnaire:

- 1) Clear, precise, and detailed measurement of the dimensional structure of managerial work.
- 2) Enhanced user acceptance. Managerial job analysis techniques tend to fact stricter requirements for face validity, acceptability and efficiency than across-occupation approaches like the CMQ. Managerial job incumbents rarely have either the time or patience for lengthy, interview-based techniques, or techniques with low face validity. Written at a tenth-grade reading level, the structured questionnaire format of EXCEL allows managerial job incumbents to easily complete EXCEL as their schedule allows. In the preliminary field test, over three quarters of managers of all levels reported that EXCEL was of above average usefulness in describing the managerial work involved in their jobs. The reported average time to complete the questionnaire was two to three hours, and the reported difficulty of completion was average. Using the same "endorse-only" strategy employed by the CMQ, I believe we will be able to significantly reduce the administration times required by EXCEL.
- 3) State of the art development standards. In the next round of field testing, we plan to use both traditional and item response theory (IRT) procedures in item retention and reliability estimation. IRT-based item bias examinations will ensure the retention of fair and unbiased items.
- 4) The production of developmental job and job level descriptions. EXCEL is primarily designed to serve as both a communication and developmental tool for managers, supervisors, and executives. By using EXCEL, organizations can collect a wide variety of information about the managerial work activities of individuals, of functional groups of managers, or of different levels of managers. For example, organizations can learn about the rate of delegation across different managerial levels, or learn what different managerial groups view as critical work activities. For management training, development, and performance appraisal, this kind of descriptive information is invaluable.

- 5) Capturing the varied and changing nature of managerial work. It is time to bring the context of managerial work to the analysis of managerial jobs. Most managerial job analysis instruments, EXCEL included, simply do not capture the varied and the continually changing nature of managerial jobs. In addition to identifying the basic functions of any given managerial job, we need to be able to get at the context -- the special demands, constraints, and choices -- that sets managerial work apart from the work of other occupational groups. Using a matrix-type format similar to the CMQ, and drawing on recent work in the field of management work theory, EXCEL is currently undergoing further expansion to capture these special constructs.

In sum, the common-metric approach to job analysis is designed to collect job analysis data for a wide variety of jobs. By supplying a set of descriptors, a "common-metric," that will be applicable to jobs of different types and natures, the common-metric approach provides organizations with an empirically-based, objective methodology to compare, contrast, and order jobs. This ability to effectively classify jobs is central to a sound development and assessment of compensation programs and selection programs. Using an endorse-only, matrix type format, the two common-metric questionnaires described today are additionally capable of collecting thousands of different pieces of information about a given job, and do so in a relatively concentrated period of time. By collecting this kind of descriptive information, both common-metric questionnaires are ideally suited to the development of job descriptions, performance appraisal instruments, and the collection of content validity evidence or the development of criterion measures for selection validation investigations. The more occupation-specific questionnaires, like EXCEL for managers, supervisors, and executives, are particularly well suited to help organizations with training and development needs.

Finally, and more generally speaking, the common-metric approach is in essence, a reaction against one-shot one-time one-purpose job analysis. The proponents of common-metric job analysis view job analysis as a process that is undertaken to describe observable work activities and observable characteristics of the job environment. The descriptors use differ only in their degree of behavioral specificity. Across-occupation descriptors, like those of the CMQ, are more abstract than occupation-specific descriptors like those contained in EXCEL. The common-metric approach encompasses even more behaviorally specific descriptors, such as task descriptors, which are common to a certain function, or element descriptors, which are common to a certain task. The only requirement of common-metric descriptors is that they be verifiable and objective, describing only observable work activities or aspects of the job environment. Under the common-metric umbrella, one can begin to discern the interrelationships between elements and tasks, tasks and jobs, jobs and job families, job families and broad occupational groups. Once we can study and define these interrelationships, then the concept of a truly integrated human resource system moves within our reach.

In closing, let me mention that the field testing of the CMQ is

currently underway, and expected to continue throughout 1990, and field testing of EXCEL is tentatively scheduled for 1991. Over the course of the next two years, we expect to have more findings to report as this research unfolds; in the meantime, we are always looking for a few good field test sites. If anyone here is interested, please do not hesitate to ask me for more information.

References for this report will be made available upon request only.

Figure 1

The Dimensional Structure of the Common-Metric Questionnaire

- Factor 1: The Interpersonal Dimension, covering:  
Demanding personal situations  
Human resource responsibility  
Employee supervision  
Internal corporate relations  
External relations  
Functional level and impact of interpersonal decisions
- Factor 2: The Decision Making Dimension, covering:  
Information processing  
Financial and human resources  
Operations and production  
Planning  
Functional level and impact of decision making decisions
- Factor 3: The Mechanical, Technical, and Physical Activities Dimension, covering:  
Physical activities  
Technical activities  
Machine, equipment, tool, and vehicle operations  
Functional level and impact of mechanical, technical, and physical activities
- Factor 4: The Work Context Dimension, covering:  
Unpleasant physical environment  
Risks and hazards  
Work autonomy  
Task and skill variety  
Task significance  
Feedback  
Task identity  
Working apparel  
Work schedule  
Reward system  
Licensing or certification  
Training, education, and experience  
Self-development

Figure 2

An Integrated Managerial Factor Structure: Seven Work-Oriented Dimensions (taken from Lozada-Larsen, 1988, 1989).

-----  
COMMON DIMENSIONS OF MANAGERIAL WORK  
-----

**Human Resources: Employee Management and Personnel Administration,** covering: Employee supervision and leadership, assessment of employee job performance, instruction and training, interviewing, hiring, and managing changes in the work force, human resource guidelines and policies, personnel programs

**Financial Resources and Fixed Assets,** covering: Budgeting, finance and cash management, purchasing, management of fixed assets, financial records and reports, financial management policies and guidelines

**Executive Long-Range Corporate Planning,** covering: Current operations planning, future operations planning, corporate goal setting, general corporate planning

**Technical Services,** covering: Research on product and service ideas, development of products and services, data processing, technical support, management of the production process, safety management, contracting out production

**Customer Services,** covering: Marketing research, advertising and promotion, sales and briefings, product and service delivery, servicing customer relations, customer relations administration

**External Relations,** covering: Legal and government relations, including laws, regulations, politics, and litigation, media relations, community affairs, contributions, professional and industrial relations, business relations, external relations policies and guidelines

**Internal Corporate Affairs,** covering: Boundary management, labor relations, informal communication, internal meetings, committees, task forces, internal consultation, internal communication and coordination policies  
-----



TRAINING ASSESSORS TO PRODUCE HIGH INTERRATER RELIABILITY  
IN EVALUATING COMPLEX WRITING SAMPLES

JADE KUAN HOFFMAN  
Personnel Selection Branch  
Personnel Commission  
Los Angeles Unified School District

CALVIN C. HOFFMAN  
Human Resources Department  
Southern California Gas Company  
Los Angeles, California

An assessor training approach was developed which produced an interrater reliability (Pearson  $r$ ) of .94 for evaluating a complex writing sample (before discussion) using assessors who were unfamiliar with job content. In addition, the interrater reliabilities of the four dimensions rated (before discussion) ranged from .84 to .92. The training approach is described and results of statistical analyses are provided.

The low inter. ter reliability of scoring writing samples has been a concern of personnel selection specialists. The variety of acceptable forms of writing styles is one possible contribution to low interrater reliability. Additionally, in some public agencies, the large number of candidates processed, difficulty in obtaining enough qualified assessors who are familiar with job content, limitations in time allotted for assessor training, and the use of complex exercises all contribute to low interrater reliability.

In most administrations, candidates are given ample time to study test materials and produce a writing sample. Assessors are typically provided with a set of test materials to review; are briefed on the established rating standards; and then, are expected to evaluate candidates' products often without time to truly study and understand the writing task.

In spite of problems associated with direct evaluation of writing skills, it is preferable to measure writing skills by writing samples rather than by objective multiple-choice methods (Quellmalz, 1986). Multiple choice methods also are more difficult to develop than are writing samples.

There are two general approaches to the scoring of writing samples, holistic and analytic (Quellmalz, 1986). In the holistic method, assessors evaluate the overall quality of a writing sample.

---

The authors gratefully acknowledge the comments of Anita Ford, T.R. Lin, Ron Marmalefsky, and Augie Ryanen.

In the analytic approach, assessors evaluate writing samples on predetermined dimensions such as spelling, punctuation, and organization to arrive at a final score.

Quellmalz (1981) favors the analytic approach, suggesting that the requirement to assign separate scores resulted in more focused ratings. The analytic approach makes it easier to provide candidates with more specific feedback regarding the basis of the assessors' ratings. In consideration of these factors, the analytic approach was deemed to be preferable for this study primarily because in the organization in question, candidates are allowed to review their rating sheets.

## Method

### Description of Job

The job in question is an entry-level technical trainee classification in a very large west coast school district. This job is one of the few technical trainee classifications which provides incumbents with an opportunity to gradually progress into a variety of higher technical or management classifications in areas of personnel, finance, engineering, school planning, and building maintenance. Although the duties and tasks performed by these incumbents vary with the assigned field, the underlying knowledge, skills, and abilities required to successfully perform these duties are very similar.

### Job Analysis

The multiple method job analysis conducted to identify relevant tasks and knowledge, skills, abilities, and personal characteristics (KSAP's) included review of written documentation, supervisory conferences, incumbent interviews, job observation, and a survey. Among all the critical KSAP's identified, written communication was one of the critical KSAP's.

Regardless of the different professional fields into which the incumbents may progress, written communication is an essential skill. Incumbents in this classification write administrative reports, project proposals, procedures, personnel selection validation reports, test materials, financial reports, etc.

### Development of Writing Assessment Material

Since a content-oriented validation strategy was used to develop the selection instruments for this classification, the writing assessment exercise was developed by simulating actual work assignments (Wernimont & Campbell, 1968). Because this is a professional classification with minimum qualifications set at the college graduate level, writing assignments also require analytical skills. Most of the writing produced by incumbents of this classification is the result of information collection and analysis. e.g., executive summary-type reports are frequently assigned to incumbents. It is usually the incumbents' responsibility to

determine which information is critical and should be included in the report as well as determining which data is not sufficiently critical to be included. Another example is that of incumbents working in the personnel selection field. The examination documentation they produce frequently summarizes data collected from literature reviews, conferences, interviews, and numerical analyses of data.

Because many of the writing assignments revolve around report writing and information analysis, these two functions were incorporated into the scenario for the writing assessment exercise. The writing exercise was designed to emphasize the candidates' ability to organize effectively and to communicate their thoughts in a clear and concise manner. A maximum time limit of 2 1/4 hours was allotted to the candidates for completion of the exercise.

Candidates were asked to assume that they were newly selected technical trainees for a school district. Their supervisor asked them to attend a meeting at which safety concerns about a recent earthquake-damaged building were discussed. The purpose of the meeting was to detail how weekend coverage was to be handled during the next several weeks, and how minimum computer coverage was to be maintained. Candidates were further told that their assignment was to take notes of what was discussed and then to write these into a narrative report to a Director who was ultimately responsible for the project but was unable to attend the meeting.

Candidates were provided with a 29-item information sheet constituting their "notes" from the meeting to be used as the basis of their reports. Appendix A lists five of the 29 items. Additionally, they were informed that the content, rather than the length, of the report was more important. It was permissible to create supporting information necessary to construct a summary of the meeting, or to make reasonable assumptions in order to fill in gaps in information. However, facts which could not be clearly derived from the notes provided were not to be included.

### Scoring Guide

An analytic scoring guide similar to the one developed by Hoffman and Holden (1990) was developed for four dimensions. Each dimension was operationally defined so that assessors could better understand what they were being asked to rate.

The four dimensions included: A) Analytical Skills, B) Spelling, C) Punctuation and Grammar, and D) Organization and Clarity. Appendix B presents the operational definition of Dimension D, organization and clarity, along with anchors and scale points for each of the four scale categories.

Of the four dimensions, Dimension A (Analytical Skills) and Dimension D (Organization and Clarity) were assigned twice as many points as Dimension B (Spelling) and Dimension C (Punctuation and Grammar). The difference in the weights reflects criticality ratings from the job analysis which will not be discussed in this paper. The maximum total score for the writing exercise was 18 points (six points for Dimensions A and D and three points for

Dimensions B and C).

### Assessor Training

During traditional orientation sessions conducted by many public agencies, assessors are given a set of test materials to review, briefed on the established rating standards and are then expected to evaluate candidates' writing products. These orientation sessions usually last about one hour. In some organizations, assessors are sent the test materials to review prior to the briefing session. This can be challenged as a procedure which has great potential for impacting test confidentiality. Due to the complexity of the writing exercise and the diversity of candidates' responses, the results of brief orientation procedures often include lower interrater reliabilities and more lengthy discussions than does the approach described in this paper.

The training approach conducted for this project differs from procedures described above; it was more similar to assessor training procedures used by many assessment centers. Instead of beginning the session by training the two assessors to assess, they were first asked to assume the role of a candidate. That is, the assessors were asked to read the instructions and other information given to candidates and then to analyze the data as if they were preparing to write the report themselves.

After carefully studying the test material, assessors were asked to independently rate the criticality of each item in the 29-item information sheet using a five-point scale. A "five" on the scale indicates the information is essential and must be provided to the Director, or else an inaccurate or inappropriate decision may be made. A "three" indicates that although the absence of this information may impact the decision made by the Director, no serious negative consequences will result from the decision. A "one" on the scale indicates the inclusion of the information may mislead the Director or is a waste of the Director's time. When each assessor had rated all 29 items, a discussion was conducted to reach agreement on the importance of each item. This procedure was designed to aid assessors in rating the analytical dimension. It would be possible to collect ratings on these items as part of the test development procedure, then use them as predetermined rating standards during assessor training. Due to the frequent administration of the examination in this organization, new test materials are developed constantly. Time constraints and test confidentiality concerns render the collection of this information prior to the administration of the test impractical in most cases.

In addition to discussing the 29-item list, another procedure was implemented to aid assessors in rating the dimensions of spelling and punctuation and grammar. Whenever an assessor located a spelling, punctuation or grammatical error, s/he marked it with a colored pencil. If the second assessor located any additional mistakes, s/he would mark it as well. This process helped to insure that assessors would base their evaluations on similar error counts.

Once all the items on the information sheet, dimension

definitions, and anchors were thoroughly discussed, the assessors were briefed on the content of the job in question, and were reminded that handwriting was not an element to be assessed.

### Subjects

The subjects consisted of 21 candidates who had successfully passed the first of three multiple-hurdle test parts, a 150-item multiple choice test. A total sample of 149 applicants took the multiple choice test. It should be noted that several dimensions measured in the writing exercise were also measured in the paper and pencil test using different content and format. Results of statistical analyses comparing the paper and pencil and writing sample tests will be discussed later in this paper.

### Procedure

The content, time allowed, and procedure pertaining to the writing sample were detailed earlier in the "Development of Writing Assessment" section. After all subjects completed the writing exercise, two assessors independently read and evaluated their products. Scores produced by the assessors were then recorded on a "Writing Project Discussion Form" (see Appendix C). This form was used to record the ratings assigned for the candidate by each assessor. Even though both assessors understood that identical scores were not required, they were instructed to discuss any discrepancy of one point or more in dimension rating. In discussing ratings, assessors were instructed to refer back to the report itself to show exactly why a specific rating was assigned and to use specific examples from the report to guide their evaluations. After the discussion, final dimension scores within one point difference were allowed.

### Results

TABLE 1  
Interrater Reliability of Dimension Ratings  
Before and After Discussion

Dimensions	Analytical Skills	Spelling	Punctuation and Grammar	Organization and Clarity	Sum
Before Discussion	.89**	.84**	.87**	.92**	.94**
After Discussion	1.00**	1.00**	1.00**	.99**	1.00**
n = 21    * p < .01 one-tailed    ** p < .001 one-tailed					



The interrater reliability (Pearson  $r$ ) of the sum of the four dimensions before discussion was .94, while interrater reliability after discussion was 1.00. In addition, the interrater reliabilities of the four dimensions rated (before discussion) ranged from .84 to .92, while reliabilities after discussion ranged from .99 to 1.00. Table 1 presents the interrater reliability (Pearson  $r$ ) of each dimension and the sum of dimension scores before and after discussion.

Table 2 presents intercorrelations among the dimensions and the sum of dimension scores. The upper diagonal of the matrix reports Pearson  $r$  after discussion and the lower diagonal reports Pearson  $r$  before discussion.

TABLE 2  
Intercorrelations of Dimension Ratings  
Before and After Discussion

	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>
1. Analytical Skills	---	.22	.22	.49	.71**
2. Spelling	.13	---	.66**	.56**	.70**
3. Punctuation and Grammar	.19	.62*	---	.73**	.78**
4. Organization and Clarity	.38	.49	.73**	---	.90**
5. SUM	.68**	.62*	.78**	.89**	---
<u>n</u> = 21    * <u>p</u> < .01 one-tailed    ** <u>p</u> < .001 one-tailed					

TABLE 3  
Descriptive Statistics on the Multiple Choice Test  
and Writing Sample \*

Variable	Mean	SD	N
1. Multiple Choice (total sample)	53.69	13.71	149
2. Multiple Choice (restricted sample)	84.81	9.08	21
3. Writing Sample (restricted sample)	8.54	3.27	21

\* Note: - correlation between multiple choice test and writing sample (restricted samples) is -.04 ns.

One additional goal of this research was to evaluate the extent to which a multiple choice measure of writing and other skills was related to performance on the writing exercise. Unfortunately,



extreme range restriction caused by using the multiple choice test as a pass/fail hurdle prior to the writing exercise made correlational analyses both misleading and hard to interpret. Table 3 provides descriptive statistics on the multiple choice test and writing sample.

### Discussion

The results of this study clearly demonstrate that high interrater reliability can be obtained even when the exercise contains complex text and the assessors are not familiar with the content of the job. Obtaining a majority of assessors with previous job experience in the same job or even in a similar field would be nearly impossible in this organization. The authors believe the training approach described in this paper provides solutions to some of these limitations.

The present training procedure forces assessors to understand the test conditions and materials set forward for the candidates. By emphasizing the understanding of the complex content of the writing exercise during the training session, disagreements in interpretation are aired and corrected prior to the actual evaluation of candidates' writing samples. This is evidenced by the high interrater reliability obtained prior to discussion. Additionally, this procedure reduces lengthy discussions to a simple reminder for the co-assessor that he/she has miscounted spelling or grammar errors or has omitted critical information. The investment of spending more time in assessor training pays off by reducing time needed for assessor discussion.

While it could be argued that the inclusion of analytical skills might contaminate the evaluation of writing skills, the results of this procedure clearly indicate that assessors could rate these skills fairly independently. A review of Table 2 indicates that analytical skills does not correlate significantly with any dimension. It does correlate significantly with total score which is analagous to an item-total correlation. This result should be expected since only four dimensions are summed to obtain the total score.

The results obtained by using this assessor training approach with the direct writing assessment method are very encouraging, although further research with larger samples should be conducted. It is recommended that organizations facing similar constraints consider using this approach to train writing skills assessors.

### REFERENCES

- Hoffman, C. C. & Holden, L. M. (1990). Development and operational use of an analytic writing evaluation instrument. Presented as part of the symposium Alternate Strategies for Assessing Writing Skills presented at the 1990 International Personnel Management Association Assessment Council Conference, June 1990.

- Quellmalz, E. S. (1981). Report on Conejo Valley's fourth-grade writing assessment. Los Angeles, CA: Center for the Study of Evaluation, University of California.
- Quellmalz, E. S. (1986). Writing skills assessment. In R. A. Berk (ed.), Performance assessment. Baltimore, MD: The Johns Hopkins University Press.
- Wernimont, P. F. & Campbell, J. P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

#### Appendix A

##### Sample Items Used in the Candidate Information Sheet

1. Mtg. on 11/1/89 w/Joe Boss, Business Manager, Carla Danish, Asst. Dir., and Peter C. Disk, Computer Operations Supvr.
2. Mr. Disk noted that salary warrants go out 3rd Wednesday of each month, addtl. staff usually required for processing the weekend before.
3. Mgrs. may consult staff for input re: schedules or may decide on schedules w/out their input.
4. Minimum staff needed: 4 per shift, except on weekend prior to warrants, where full staff requested, but could "get by" with 8-8-8 Sat, 4-4-4 Sunday.
5. Mtg. lasted 2 hrs. - concluded at 3:00 pm.

#### Appendix B

##### Sample Writing Exercise Rating Standards and Rating Form

#### D: ORGANIZATION AND CLARITY

Are thoughts and ideas clearly expressed or is there doubt or confusion as to the writer's intention? Is the report understandable?

6-5: Report was very well organized and clearly written; easily understood.

4-3: Report was well organized with a few exceptions; mostly clear and easily understood.

2-1: Report had some flaws in organization and clarity; some of the intended message was provided; but some would not be clearly understood.

0: Report was very disorganized and unclear; would not provide needed information to others.

## Appendix C

### Writing Project Discussion Form

Candidates's I.D. #: \_\_\_\_\_ Assessor 1 ID #: \_\_\_\_\_ Assessor 2 ID #: \_\_\_\_\_

Record the ratings assigned for the candidate by each assessor. For any discrepancy of one point or more in individual ratings on a dimension, discuss your ratings and reach final dimension scores within one point of each other. During the discussion, refer to the exercise itself to show exactly why a specific rating was assigned. Use specific examples from the report to guide your evaluation.

	Dimension	Before Discussion		After Discussion	
		Assessor 1	Assessor 2	Assessor 1	Assessor 2
A:	ANALYTICAL SKILL	_____	_____	_____	_____
B:	SPELLING	_____	_____	_____	_____
C:	PUNCTUATION AND GRAMMAR	_____	_____	_____	_____
D:	ORGANIZATION AND CLARITY	_____	_____	_____	_____
<u>TOTAL SCORE</u>		_____	_____	_____	_____

THE "NESTOR FACTOR":  
MEASURING QUALITY IN THE SCIENCE AND  
ENGINEERING WORKFORCE

FOR PRESENTATION AT THE SYMPOSIUM ON  
ASSESSMENT OF WORKFORCE QUALITY AND EMPLOYABILITY SKILLS  
1990 INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION CONFERENCE  
JUNE 25, 1990

JEANNE CARNEY  
STAFF SPECIALIST  
DEPARTMENT OF DEFENSE

## INTRODUCTION

Nestor is the loquacious old warrior of the Iliad who continually exhorts the Achaians to battle against the Trojans. To energize the troops, he is especially fond of recounting stories of his generation's daring deeds and of comparing the great warriors of his youth with the "youngsters" who accompanied Agamemnon and Menelaos to Troy. Of course, he always intimates that the current generation doesn't quite measure up. His thesis throughout the Iliad is essentially, "They just don't make heroes like they used to."

For several years, I've been hearing somewhat the same thing about the federal workforce. Put more explicitly, the lament is that the best and the brightest are leaving, and the quality of those whom we are hiring to replace those leaving is declining. These assertions were first made about the science and engineering workforce; more recently, they are being applied to the federal workforce as a whole.

But just like Nestor, for many years we've had little more than anecdotal information to prove or disprove this hypothesis. Few attempts have been made to define what we mean by quality, to establish concrete measures of quality factors, or to assess the presence of these factors in the workforce over time to determine whether they are, in fact, eroding.

## THE DoD EXPERIENCE

Of the one million civilians who work for the Department of Defense, about 100,000 or roughly 10 percent are scientists and engineers. One-quarter of these work in DoD's 70 Research and Development laboratories. The concern that we were losing our best and brightest scientists and engineers (S&E's) was first expressed by the directors of these laboratories during a study the Department conducted during 1982.

Even though DoD accession and loss data showed that recruitment and retention were in balance, (in other words, we were hiring as many laboratory S&E's as we were losing), the 1982 laboratory study showed that almost 60 percent of those we were losing were at the journeyman, or GS 12/13 grade levels, and those we were recruiting were principally at the entry-level grades. The laboratory directors asserted that among those we were losing were some of our "superstars," and although the absolute numbers were small, these S&E's made the significant scientific and engineering advances on which we depended to maintain the nation's technological advantage.

In 1986 we decided to update this early laboratory study of recruitment and retention trends among our laboratory S&E's. In our update, we found that our attrition rate had grown from 6 percent of the S&E workforce in the early 80's to 10 percent by

1986. Loss rates were up over 1982 figures for every civil service grade level except GS 5 and 7. In fact, DoD laboratories' losses in 1986 ranged from 9 to 11 percent of S&E's at the GS-12-14 levels. (A U.S. Merit Systems Protection Board Study showed 1987 losses at these grade levels to be about 5 percent federal-wide.)

Of the 2,475 S&E's who left the laboratory in FY 86:

- 26 percent were GS 12's
- 23 percent were GS 13's
- 17 percent were GS 14's, and
- 8 percent were GS 15's.

It is interesting to note that the percentage of those leaving at each grade level closely mirrored the percentage of each grade level in the laboratory population, i.e. in 1986 approximately 27 percent of DoD's 25,000 laboratory S&E's were GS-12's; 25 percent were GS-13's; 16 percent were GS-14's and 7 percent were GS-15's. Between 1978 and 1986, resignations (as opposed to retirements, transfers and other reasons for separation) climbed from 40 percent to 50 percent, and of those resigning, 62 percent departed to accept a job in industry. Salary and opportunity for advancement were cited most as reasons for resigning.

By 1986, almost three-quarters of all vacancies at the GS 9-12 levels were being filled through merit promotion, up over 10 percent from 1982 data. This suggested that it had become harder to fill these positions through new hires and through transfers from other federal agencies. It was also interesting that irrespective of growing opportunities for promotion, loss rates of S&E's at these same levels in the laboratories continued to climb as shown above.

DoD laboratories ended FY 86 with over 1,500 vacancies, 500 more than we found in 1982. Vacancies included the following in critical disciplines:

- 450 electronic/electrical engineers,
- 146 mechanical engineers,
- 156 computer scientists,
- 127 general engineers, and
- 108 aeronautical engineers.

Although these trends appeared to be somewhat disturbing, they still didn't tell us much about the quality of the people we were losing, gaining or retaining in the DoD S&E workforce. And so, we embarked on a second study, this time an attempt to collect data on the characteristics of the workforce and determine whether these characteristics were indicators of quality.

Before describing the approach we have taken in attempting to assess quality in the S&E workforce, however, I want to



recount a personal experience that is illustrative of some of the problems related to measuring quality.

Twenty-three years ago, I accepted a temporary GS-3 clerk-typist job with the federal government. This was an act of desperation: I had left a teaching job in June based on assurances from various federal sources that I would be hired quickly for any one of several professional positions for which I had applied. But when I still had no job by August due to delays in the security clearance process, I sought out a temporary clerk-typist position.

Even though I was certified as a GS-7, I was required to take additional tests to qualify at the lower GS-3 clerk-typist level. These tests included a timed typing test that was normally given only once to an applicant. You either passed or you failed. I shall be ever grateful, however, to the woman who gave it to me five times so that I could muster the 40 words per minute with 5 errors that resulted in a three-month temporary clerk-typist appointment and allowed me to return to eating regularly.

I soon learned how valuable accurate typing could be, especially since it was beyond my abilities. The office to which I was assigned required each document to be typed on impossible-to-erase watermarked bond, accompanied by 8 carbon copies. Word processors, self-correcting typewriters, and even "white-out" were things not yet even imagined. As my embarrassment grew daily over the ruined reams of letterhead in my wastepaper basket, I began to bring my mistakes home each evening in ever larger handbags.

After several weeks of agony, the fearsome head of the office glowered at me one afternoon and said in tones audible to all: "I have been watching you for some time now. It is obvious to me that you cannot type. Please give me your resume so that I can see what it is you can do."

Frightened that I would soon be fired, I did as he said. He reviewed my resume and suggested that I apply for a professional opening in the division, which I did, and for which I was eventually hired. He immediately ordered that my typewriter be taken away from me, and so began my federal career.

I tell this story not to reveal my inadequacies, but to make two points: (1) if we had been measuring the quality of clerk-typist hires in 1967, I would have been a prime example of the declining quality of entry-level clerks; and (2) if it is at all possible to measure quality, we may find that it can only be measured in relation to the requirements of the job to be performed.

## DOD STUDY OF THE CHARACTERISTICS OF THE S&E WORKFORCE

The DoD study of S&E workforce characteristics has set out to answer the following questions:

What are the characteristics of the DoD S&E workforce?  
What characteristics can be considered measures of quality?  
To what segments of this workforce do these measures apply?  
How do we know that these measures are valid?  
How can changes in time in these characteristics be shown?

To begin, we conducted a rather extensive literature search to learn whether others had undertaken similar studies of comparable populations. Not only did we want to identify measures others had found to be valid that we could adopt, but we hoped for external baseline populations against which to compare findings about our own internal workforce.

Unfortunately, while we found over 100 studies of various S&E occupations, some of the workforce descriptors examined in these studies, such as the amount of overtime worked, did not appear to be relevant measures of quality. Others, such as publications and patents did appear to be valid measures of S&E quality, although data on comparable populations were sketchy. We gave up the hope of designing a questionnaire for our sample population which would ask the identical questions asked of external S&E sample populations for purposes of baseline comparisons. Instead, while we adopted some questions, we developed many of our own which are more specific to a federal population. We designed a questionnaire totaling 53 questions to collect a broad spectrum of data ranging from the respondent's education to his or her current attitudes about working for DoD.

We selected approximately 21,000 of DoD's 100,000 S&E's to be part of the survey. These S&E's work in 89 different job series, and include occupations in the social sciences, biological and health sciences, physical sciences, mathematical sciences and all engineering disciplines. The sample was drawn to encompass sufficient numbers of each discipline to be statistically valid. It was also drawn to reflect organizational location (Army, Navy, Air Force, Defense Agencies) and type of work assignment or functional classification (research, development, test and evaluation, management, production, data collection and analysis, and operations and maintenance). Of the 21,000 S&E's sampled during the summer of 1988, over 15,000 responded to the survey. The oldest of the respondents joined DoD prior to 1958, the youngest in 1987.

In early 1989, we began analysis of the data, dividing our sample population only by entry cohort groups to see how the characteristics of the population had changed over time. Our first runs of the entire population showed some alarming trends.

When we compiled the data on educational level, we found an almost steady decline, with only a few anomalies, in the percentage of S&E's with advanced degrees in each cohort group since 1958.

Scientific papers and patents are often used as another measure of S&E quality and productivity characteristics. Again our initial analyses showed that those who joined DoD in later years produced fewer papers and patents during a recent five-year period than those who joined DoD in earlier years.

But even as we shared some of these initial findings at conferences on workforce quality during 1989, we were uncomfortable with them. Unfortunately, it took us a long time to realize why and then to decide what to do about it, but we are now revising our data analysis plans.

In the same way that my skill at typing is not a valid measure of my ability to perform the job I now hold, we have concluded that some measures of quality are valid only for certain segments of the population. Hence, the traditional metrics of patents and publications are only valid measures of quality for that fraction of the S&E workforce that engages in research and development. Advanced degrees are frequently only needed in a research setting, although many older S&E's who have since moved from hands-on R&D to management positions do hold them. Measuring the quality of an engineer conducting repairs in a Naval Air Rework Facility by his output of scientific papers would be as inappropriate as using the same measure to assess the quality of the engineer working on the floor of an automotive plant in Detroit.

Figures 1, 2 and 3 show the differences in these characteristics in the population as a whole versus only the R&D portion of our S&E sample.

In Figure 1, the percentage of S&E's with PhD's is shown for all S&E's in the sample, as well as for those who indicated their functional classification was research or development (R&D). While the period 1983-86 appears to have been a real low point in hiring PhD S&E's for both DoD as a whole and the R&D Community in particular, some recovery seems to have occurred by 1987. In fact, at least for PhD scientists, 1987 was a better year for hiring PhD's than the period 1958-1967.

**FIGURE 1 - PERCENTAGE OF S&E'S WITH PHD'S BY YEAR DOD EMPLOYMENT BEGAN**

<b>ALL S&amp;E'S IN SAMPLE POPULATION</b>						
	1987	1986-83	1982-78	1977-68	1967-58	<1958
Scientists	20.2	17.4	22.1	28.0	21.0	35.8
Engineers	1.8	1.0	3.6	5.2	4.1	7.8
<b>R&amp;D S&amp;E'S IN SAMPLE POPULATION</b>						
	1987	1986-83	1982-78	1977-68	1967-58	<1958
Scientists	30.9	22.8	31.3	42.4	27.5	44.0
Engineers	6.0	2.9	10.7	11.5	9.1	15.7

Figure 2 contrasts publication output of all S&E's in the sample with that of R&D S&E's. While the output of publications among all S&E's in the sample appears to be declining, the R&D S&E's show only a slight variation in the number of publications produced by each cohort group during the five-year period of 1983-1987. Moreover, it is interesting to note that supervisors and members of the Senior Executive Service (SES) appear to have the most publications and papers to their credit. One reason for this may be that productive people are recognized and rewarded through the merit promotion system and a high level of correlation between output and an individual's position is simply to be expected. Another reason may be that managers of scientists and engineers have the opportunity to make technical contributions to the work of their subordinates, and are, therefore, credited in the papers of their staff members.

FIGURE 2 - PUBLICATIONS PRODUCED 1983 - 1987\*

ALL S&E'S IN SAMPLE POPULATION

	1982-78	1977-68	1967-58	< 1958
Scientists	3.5	4.7	4.4	4.8
Engineers	1.9	2.3	2.7	3.1
Supervisors	3.2	3.7	3.5	3.9
SES Members	5.6	6.8	5.6	6.3

R&D S&E'S IN SAMPLE POPULATION

	1982-78	1977-68	1967-58	< 1958
Scientists	6.2	7.6	6.0	6.6
Engineers	3.8	4.0	4.1	4.5
Supervisors	6.2	7.2	5.6	7.5
SES Members	7.3	10.0	8.5	9.7

\*For purposes of this paper, S&E's in the sample who joined DoD in 1983 or later are not included in this tally.

Our first look at patent data in Figure 3 gave us cause for concern, the decline in this category for all S&E's being obvious. Moreover, the R&D S&E patent data appear to be an anomaly when compared with the publications data in Figure 3 where R&D S&E's in all cohort groups make a strong showing. One explanation may be that patent application and grant data correlate more closely with the number of years the S&E's have worked for DoD and the length of time inherent in the patent application process than with any other factor. Another explanation is that until the enactment of the Federal Technology Transfer Act of 1986 there was little incentive for federal S&E's to pursue the lengthy patent application process. This Act, by providing opportunities for federal S&E's to share in royalties, may cause the number of patents to rise in future years. Finally, DoD has been experiencing a severe shortage of patent attorneys, resulting in a backlog of patent applications averaging three to four years.

FIGURE 3 - TOTAL PATENTS APPLIED FOR OR GRANTED\*

ALL S&E'S IN SAMPLE POPULATION

	1982-78	1977-68	1967-58	< 1958
Scientists	.2	.6	1.4	1.7
Engineers	.2	.6	1.1	1.5
Supervisors	.3	.7	1.2	1.5
SES Members	1.0	2.0	1.4	1.9

R&D S&E'S IN SAMPLE POPULATION

	1982-78	1977-68	1967-58	< 1958
Scientists	.5	.9	2.4	2.6
Engineers	.5	1.2	2.2	3.1
Supervisors	.7	1.3	2.6	3.4
SES Members	2.8	3.1	4.1	3.1

S&E's in the sample who joined DoD in 1983 or later are not included in this tally.

It is clear to us now that our methodology for analyzing the data we gathered must take into account the appropriateness of these measures to the particular population, viz. some measures are appropriate only to those S&E's in the sample who are engaged in R&D. But we are still faced with two other problems: (1) how do we determine the validity of these measures when applied to the R&D S&E's in the sample and (2) if these are valid measures for R&D S&E's, what measures do we apply to everybody else?

Our current plan is to separate the R&D S&E's from the rest of the sample, and further subdivide the R&D S&E's into scientists and engineers. Next, using questionnaire responses that identify publications, patents, awards, and chairs and fellowships in professional societies, we intend to group R&D scientists and engineers into three categories: the top performers, the middle and the bottom. We've selected these factors because they are measures of excellence applied by the S&E community at large, external to the judgment of anyone within the Department. Without identifying individuals by name, we can obtain records that will allow us to determine whether the performance appraisals of these individuals correlate with their ranking at the top, middle or bottom. We intend to conduct further tests to determine whether promotion rates also



correlate with the quality characteristics of the top, middle and bottom groupings.

Our next step will be to subdivide these groupings by entry-year cohorts and determine the characteristics each group possessed upon entering the workforce. These characteristics include: schools attended; class standing; grade point average; educational level; participation in co-op or intern programs at entry; sources of graduate support; average civil service grade and step at entry; and formal offers of employment.

We then intend to identify the characteristics each group now possesses. These characteristics include: average grade and step; percent who are supervisors; percent who are members of the Senior Executive Service; promotion rate; participation in continuing education; professional status (of engineers); and teaching as a secondary activity. We also will conduct the above analyses for S&E's in such critical fields as electronic/electrical engineering, physics, and computer science, and separately for some 1,500 S&E's who participated in the survey, but have since left DoD.

Finally, we will compare each group's responses to "quality of life" questions included in the questionnaire. These include why they joined DoD; why they want to leave; whether they are seeking employment outside the federal government; and what are the most and least satisfying aspects of their jobs and working for DoD.

Based on what we learn from the above analyses, we will proceed to do the analyses on the remainder of the non-R&D S&E sample. If we find that the quality measures of publications, patents and other external assessments of quality correlate with the internal assessments of performance appraisals and promotion rates, we may then be able to use these internal assessments in sorting our non-R&D S&E's (who are unlikely to have produced publications and patents) into appropriate groupings for analysis.

## CONCLUSION

This study has taken far longer than we originally anticipated when we first set out to see whether quantitative measures could be applied to the elusive concept of quality. We have followed many false leads, tried and rejected numerous theories and the study designs that were based on them, and have made countless mistakes. Mostly, we have learned what not to do, more than what we should do, in attempting to measure quality. Old-timers might attribute our fumbblings to a lack of quality in the study sponsor and study team. But, then again, as far as we've been able to learn, few old-timers ever even tried to find the facts and prove or disprove the "Nestor Factor."

## OPM'S QUALITY ASSESSMENT PROGRAM

Jay A. Gandy  
U.S. Office of Personnel Management

In response to calls for information and action to assess and improve Federal workforce quality, OPM's Office of Personnel Research and Development (PRD) has formulated and is implementing a broad-based program of quality research and development. This program has been identified as one of OPM's critical milestones under the President's Management By Objectives System.

The PRD program recognizes the complexity of quality measurement as a concept which goes beyond individual attributes to encompass measures of organizational effectiveness and client satisfaction.

Mindful of the complexity of quality issues, OPM and the Merit Systems Protection Board (MSPB) have jointly created an Advisory Committee on Federal Workforce Quality Assessment to support both OPM and MSPB efforts to collect reliable data on workforce quality. The Committee, consisting of 25 individuals drawn from Federal, State, and local governments, private industry, academia, unions, and professional associations, will review workforce quality assessment projects, review data and advise on its analysis and interpretation, advise on appropriate responses to research findings, and help coordinate Federal and non-Federal efforts. Chartered for two years, the first meeting of the committee was held May 1, 1990.

OPM has initiated a number of projects to collect data. In general, they can be grouped into the following areas: (1) collection of applicant and new hire data over time; (2) special studies to collect data on selected critical occupations; and (3) collection of non-Federal comparison data.

Since May 1989, applicant quality data have been collected for five occupations -- clerical, contract specialist, computer specialist, internal revenue officer, and accountant/auditor. Coverage is being expanded to additional occupations, including most of those with selection through OPM-administered automated examinations, as well as to applicants and new hires under agency direct hire and delegated examining authorities. The data collection will include traditional measures of quality, such as academic achievement and will be supplemented with information from automated records, such as test scores. We will track the quality of applicants, new hires, and incumbents over time, to identify trends in the Federal workforce, and to establish whether quality is changing. This data collection will also facilitate comparisons of Federal quality with the quality of comparable segments of the non-Federal labor force.

The absence of standardized measures is a serious limitation

in workforce quality assessment. We have begun a project to develop a group administered adult literacy assessment instrument. If the project goes according to plan, specifications and norms for this instrument will be linked to the Department of Education's National Adult Literacy Survey. Linkage is also planned with parts of the Department of Defense Armed Services Aptitude Battery (ASVAB). These linkages will enable comparisons to be made of civilian and military applicants and hires with representative samples in the national population.

While workforce quality is a general concern for the entire Federal workforce, several critical occupations will be targeted for in-depth analysis each year. These are occupations for which agencies and OPM have identified special recruitment and retention concerns. Data will be collected during FY90 from a nationwide sample of incumbents in two occupational groupings -- scientist/engineer and computer specialist; a third study will begin for economist/statistician. These surveys will cover (1) education, experience, and achievement, (2) job satisfaction, and (3) job performance. The incumbent surveys will be supplemented by a specially designed supervisory performance appraisal. These incumbent studies will assist in developing a broad base of information on quality indicators, establish baseline information to track quality changes over time, and facilitate comparison with similar jobs in non-Federal organizations.

Because the Federal government, as an employer, competes with other employers, Federal quality cannot be considered in isolation. Accordingly, one objective of the OPM program is to relate Federal quality to comparable groups of employees in private organizations. During FY90, a study of private sector "anchor" data will be collected from existing secondary sources, e.g., Census Bureau and Bureau of Labor Statistics, and from professional associations. A pilot project is also underway, with the assistance of the Society for Human Resource Management, to collect comparison data directly from private companies.

The OPM quality program includes development of an information base on current and projected skill deficits in the Federal workforce. To assist in formulating development strategies, a nationwide sample of senior Federal managers will be surveyed during fiscal year 1991 to determine how they view changes taking place in the Federal workforce. The data collected in the survey will be used to develop a government-wide profile of Federal workforce skills needs.

An important aspect of OPM's quality program is the collection and distribution of information about research findings and projects to maintain and upgrade Federal workforce quality. To facilitate this, OPM has established a government-wide Technical Contact Network. An Annual Federal Workforce Quality Summary is also planned, as are focus groups and technical workshops.

# **MICHIGAN'S APPROACH TO ASSESSING EMPLOYABILITY SKILLS<sup>1</sup>**

by Paul M. Stemmer, Jr., Ph.D. and William L. Brown

Michigan Department of Education  
Office of Technical Assistance and Evaluation  
PO Box 30008  
Lansing, Michigan 48909

Many people are familiar with the changes in our society, both in terms of the changing workforce literacy requirements by industry and the increasing failure of our institutions to keep up with this pace. Most studies of these new requirements focus on academic literacy demands. The Michigan Employability Skills Task Force has suggested that Personal Management and Teamwork skills are also important. They further contend that much of the reluctance to address Personal Management and Teamwork skills is related to our difficulty in measuring these skills validly and reliably. The results of a state-wide survey have confirmed that the task force is on the right track. This paper will discuss the skills we have found important and consider alternative ways of assessing these skills and then suggest a way of linking assessment results to job opportunity.

## **A Brief History of the Michigan Employability Skills Program**

In July, 1987, the State of Michigan Legislature adopted an Employability Skills Assessment program in the schools. In November of that same year, Governor Blanchard's Commission on Jobs and Economic Development co-chaired by Lee Iacocca (President of Chrysler Motors Corp.) and Douglas Fraser (former President of the UAW) convened the Employability Skills Task Force and charged it with the responsibility for identifying the generic skills and behaviors employers believed to be important across a broad range of business, service and industrial sector jobs.

The Task Force was co-chaired by Peter J. Pestillo (Vice-President, Ford Motor Co.) and by Stephen Yokich (Vice-President, UAW). The Task force represents a broad spectrum of business, labor, government and education leaders.

The Governor's Cabinet Council on Human Investment directed by Gary Bachula, devised a strategic policy to create one integrated Human Investment System out of the state's departments, agencies and programs with the common purpose of better accountability and management of Michigan's wide array of human investment resources (see **Creating a Human Investment**

---

<sup>1</sup>Presented at the 1990 Annual Conference of the International Personnel Management Association, Assessing Workforce 2000: San Diego, California. The views, opinions and findings in this report are those of the authors and do not reflect the official position of the Michigan Department of Education.



**System: Report to the Governor)<sup>2</sup>.** The Human Investment Fund Board was charged to:

- Develop and oversee an integrated policy strategy for managing and expending Michigan's human investment resources across all suitable departments, agencies, and programs at state, regional and local levels.
- Establish all necessary standards needed to effectively provide comprehensive human investment services to Michigan citizens.
- Establish clearly defined outcome measures applicable across all state and federally funded human investment programs, providing consistent cross-program evaluations of results and costs.
- Advise the Governor on the implementation of the Michigan Opportunity Card (an electronic "smart card" for storing personal data for Michigan citizens).
- Work closely with the private sector to ensure that state human investment strategies developed by the Board are responsive to the needs of Michigan employers and workers.

### **Findings from the Employability Skills Survey<sup>3</sup>**

#### *Background*

#### Description of the Sample

The survey consisted of a list of 86 specific skills and behaviors arranged in the same order as the Task Force Profile (see Appendix A). Respondents rated each skill using the following statement: In my business, I need employees who can..... The responses used the following Likert Scale (with corresponding codes):

CRITICAL	HIGHLY NEEDED	SOMEWHAT NEEDED	NOT NEEDED
code: 1	2	3	4

About 7,500 surveys were mailed to employers across the State of Michigan using a mailing list supplied by the Michigan Employment Security Commission (MESC). This sample was drawn from the entire universe of employers since all are required to file with the MESC. The sample was stratified according to the type of industry using the U.S. Dept. of Labor's Standard Industry Codes, and the size of business are listed in Table 1.

---

<sup>2</sup>Creating a Human Investment System: Report to the Governor. (1989). The Michigan Job Training Coordinating Council, Box 300039, Lansing, MI 48909

<sup>3</sup>A comprehensive interpretation of the survey results can be found in Mehrens, William (1989). **Michigan Employability Skills Technical Report**. Available by writing to Dr. Mehrens at Michigan State University, 462 Erickson Hall, East Lansing, MI 48824-1034.

## Assessing Employability Skills

The eventual return rate was 2752 or about 37% of the sample. The breakdown of responses are listed in Table 1.

TABLE 1  
Employer and Sample Proportions

CATEGORY	NUMBER EMPLOYED	TOTAL GOODS	TOTAL SERVICES	SAMPLE GOODS	SAMPLE SERVICES	RETURN GOODS	RETURN SERVICES
1	1-4	11,358 8.1%	54,925 39.6%	209 2.5%	949 11.1%	53 2.0%	209 7.7%
2	5-9	5,729 4.1%	24,055 17.3%	260 3.1%	1,015 11.9%	83 3.1%	318 11.7%
3	10-19	4,466 3.2%	14,848 10.7%	331 3.9%	1,048 12.3%	126 4.6%	325 12.0%
4	20-49	3,519 2.5%	9,607 6.9%	385 4.6%	1,020 12.0%	147 5.4%	316 11.7%
5	50-99	1,383 1%	3,744 2.6%	313 3.7%	787 9.2%	104 3.8%	253 9.3%
6	100-249	910 .7%	2,229 1.6%	241 2.8%	577 6.8%	81 3.0%	209 7.7%
7	250-499	331 .2%	809 .4%	181 2.1%	406 4.8%	66 2.4%	143 5.7%
8	500-999	140 .1%	339 .24%	132 1.6%	346 4.1%	43 1.6%	135 5.0%
9	1000 +	119 .08%	218 .16%	120 1.4%	217 2.5%	40 1.5%	90 3.3%
Total		27,955 20%	110,774 80%	2,172 25%	6,365 75%	743 27%	2,009 73%

Cell frequencies represent actual count; percentages are based on column fraction of total count for that cell.

## Results

### Overall Results

Recall that 1 is critical, 2=highly needed, 3=somewhat needed, 4=not needed. As one can readily see from Table 2, almost all the 86 skills were rated to some extent within a "needed category." The overall grand mean was 2.1 (Highly Needed). Forty percent of the skills received mean responses higher than the category of Highly Needed and 76% of the skills received ratings above the mid-point on the scale. Table 2 represents an univariate description of the results; the survey items reflect the question as it was worded on the survey.



# Assessing Employability Skills

TABLE 2.  
Univariate Results by Item  
(Percent Responding by Item & by Response Category)

SURVEY ITEM NO.	MEAN*	CRITICAL	Percent of Respondents		
			HIGHLY NEEDED	SOME-WHAT NEEDED	NOT NEEDED
ACADEMIC (Questions 1 - 47)					
Q1. Pay attention to the person speaking	1.4	62	33	3.0	0.1
Q2. Ask questions to clarify understanding	1.5	53	39	6.0	0.5
Q3. Follow directions given verbally	1.4	61	34	3.0	0.0
Q4. Answer questions accurately	1.6	45	42	9.0	1.0
Q5. Explain ideas to others	2.1	22	46	27	3.0
Q6. Understand a foreign language	3.9	0.6	0.5	9.0	8
Q7. Recognize and use specific company and business terms	2.2	21	39	32	6.0
Q8. Recognize and understand signs and symbols in the workplace	2.1	29	34	26	9.0
Q9. Recognize and understand enough words to read simple instructions	1.6	53	34	9.0	1.0
Q10. Recognize and understand enough words to read complex instructions	2.1	32	35	23	8.0
Q11. Follow written instructions required for new tasks	2.0	31	40	19	6.0
Q12. Understand and evaluate written materials	2.1	27	40	24	7.0
Q13. Know how to read and use different kinds of written materials at work (e.g., letters, memos)	2.1	28	35	25	9.0
Q14. Combine and use information from several different sources	2.3	21	35	28	11
Q15. Write legibly	2.0	31	43	21	3.0
Q16. Spell correctly	2.1	25	39	27	6.0
Q17. Write sentences and paragraphs using correct punctuation and grammar	2.5	19	27	32	19
Q18. Organize and translate written thoughts into written communication	2.5	18	29	31	18
Q19. Use writing as a normal part of the job (e.g. messages, job orders, notes)	2.2	27	35	25	11
Q20. Perform basic calculations (addition, subtraction, multiplications, division)	1.8	45	34	16	4.0
Q21. Perform calculations involving fractions/percent/decimals	2.4	23	28	30	17
Q22. Read and understand diagrams, charts, graphs, and tables	2.7	14	26	35	23
Q23. Measure using U.S. measuring system	2.4	23	26	30	18
Q24. Measure using metric measuring system	3.3	7.0	11	27	53
Q25. Calculate distance, weight, area, volume and time	3.0	10	18	31	38
Q26. Understand and apply simple probability and statistics (e.g. calculate averages)	3.2	5.0	15	32	46
Q27. Estimate numerical results and judge accuracy	2.8	12	24	30	32
Q28. Estimate cost, time, and materials necessary to complete a task	2.7	14	24	32	26
Q29. Know where and how to get specialized information	2.2	25	37	28	9.0
Q30. Distinguish between fact and fiction	2.3	23	37	29	9.0
Q31. Adapt work skills to new technology	2.3	20	38	28	12
Q32. Understand and use computer/data processing terminology	2.9	12	24	26	35
Q33. Apply basic knowledge of natural science (i.e. biology, chemistry, physics)	3.3	6.0	10	27	53
Q34. Apply basic knowledge of social sciences arts, and humanities	3.2	4.0	15	32	45

# Assessing Employability Skills

Table 2 continued.  
Univariate Results by Item  
(Percent Responding by Item & by Response Category)

SURVEY ITEM NO.	MEAN*	CRITICAL	Percent of Respondents		
			HIGHLY NEEDED	SOME- WHAT NEEDED	NOT NEEDED
Q35. Operate technical equipment, instruments, and tools (e.g. gauges, meters, scales)	2.6	23	19	27	29
Q36. Determine the right tool for a task	2.5	23	26	23	25
Q37. Follow safety rules for specific equipment	2.1	39	23	18	17
Q38. Use calculators to solve problems	2.5	21	27	30	20
Q39. Demonstrate keyboarding skills	2.9	13	21	25	39
Q40. Use computer/data processing applications (e.g. word processing, business applications)	2.9	15	19	23	42
Q41. Demonstrate computer programming, networking, design and manufacture skills	3.4	4.0	11	24	59
Q42. Recognize and define problems on the job	1.8	37	43	16	3.0
Q43. Describe problems in the operation of equipment or in processes	2.3	25	39	25	13
Q44. Analyze problems to determine their sources and importance	2.2	23	40	25	8.0
Q45. Develop and evaluate new approaches to problem solving	2.4	16	37	32	14
Q46. Select the best solution to a problem	2.1	28	41	21	7.0
Q47. Carry out a decision and evaluate its effectiveness	2.2	21	41	27	9.0
PERSONAL MANAGEMENT (Questions 48 - 73)					
Q48. Apply knowledge of one's personality traits (e.g. interests, values, strengths, weaknesses) when setting personal goals	2.5	11	38	37	12
Q49. Follow a plan to achieve career goals	2.7	9	32	39	17
Q50. Exhibit self-esteem and self-confidence	1.9	27	55	14	2.0
Q51. Exhibit skills that apply to more than one job	2.0	23	50	19	4.0
Q52. Understand employees' legal rights and responsibilities	2.5	13	31	41	13
Q53. Pursue personal goals that support the organization's goals	2.3	15	44	29	29
Q54. Show respect for others	1.4	62	33	3.0	0.0
Q55. Show pride in one's work	1.4	61	35	2.0	0.3
Q56. Show enthusiasm for the work to be done	1.5	53	41	4.0	0.2
Q57. Demonstrate honesty and integrity	1.2	77	20	0.9	0.0
Q58. Meet requirement for punctuality and attendance	1.4	62	34	2.0	0.2
Q59. Plan and organize to complete tasks	1.7	40	47	9.0	0.9
Q60. Show initiative; be a "self-starter"	1.6	44	47	6.0	0.4
Q61. Meet or exceed requirements for work quality	1.6	43	51	4.0	0.1
Q62. Complete tasks in the face of job pressures and stresses	1.6	46	45	6.0	0.7
Q63. Follow safety rules and practices (e.g. hazardous materials or machinery regulations)	1.9	45	28	12	13
Q64. Demonstrate self-control	1.6	49	42	6.0	0.1
Q65. Demonstrate appropriate grooming and dress, and practice good personal hygiene	1.7	43	43	11	0.9
Q66. Be free from substance abuse (i.e. dependence on alcohol or drugs)	1.2	81	15	1.0	0.4
Q67. Work productively with minimum supervision	1.5	50	45	2.0	0.0
Q68. Know what is necessary to upgrade one's knowledge and skills	2.0	22	56	20	2.0
Q69. Demonstrate a positive attitude toward learning and growth	1.8	32	57	8.0	0.5

## Assessing Employability Skills

Table 2 continued.  
Univariate Results by Item  
(Percent Responding by Item & by Response Category)

SURVEY ITEM NO.	MEAN*	CRITICAL	Percent of Respondents		
			HIGHLY NEEDED	SOME- WHAT NEEDED	NOT NEEDED
Q70. Be adaptable, flexible and open to change	1.7	36	53	8.0	0.3
Q71. Participate in education and training	2.1	25	46	23	4.0
Q72. Use creativity and imagination on the job	2.1	25	45	25	2.0
Q73. Generate new ideas for getting a job done	2.1	22	49	24	2.0
TEAMWORK (Questions 74 - 86)					
Q74. Accept organization's mission and goals	1.8	36	50	12	1.0
Q75. Represent organization in a positive manner	1.6	44	46	8.0	0.4
Q76. Follow organizations rules, procedures, and policies	1.6	43	48	7.0	0.5
Q77. Show interest in organizations future	1.8	33	52	12	0.7
Q78. Use a team approach to identify problems and devise solutions	2.0	27	49	19	3.0
Q79. Communicate effectively with all members of the work team	1.7	43	47	8.0	0.9
Q80. Compromise to achieve work team results	1.9	28	51	1	2.0
Q81. Show sensitivity to the thoughts and opinions of others in the work team	1.8	32	53	11	1.0
Q82. Accept decisions made by work team	1.8	31	54	11	1.0
Q83. Determine when to be a leader or a follower	2.1	22	50	22	4.0
Q84. Cooperate with others to get the job done	1.6	42	49	6.0	0.4
Q85. Accept constructive criticism of performance and ideas	1.7	36	54	8.0	0.4
Q86. Show sensitivity to the needs of women and ethnic or racial minorities	2.1	29	40	22	6.0

\*MEAN is computed by averaging the following codes:

1 = Critical                      3 = Somewhat Needed  
2 = Highly Needed            4 = Not Needed

### *The Skills Ranked as Most Critical*

Table 3a shows the seven most critical skills according to their average (mean) scores. It was not surprising to us that being free from substance abuse was the most critical skill. The most surprising aspect of this table is that 5 out of 7 (71%) skills were in Personal Management skill area. Much of these skills deal with honesty, being punctual, pride and respect, etc. Many of these skills are not emphasized in traditional education and training programs. It raises the question about whether these skills were traditionally skills that were taught by parents in the home. Has the home environment abandoned these skills? Should the schools and training programs be teaching these skills?<sup>4</sup>

<sup>4</sup>There has been some argument whether these skills should more appropriately be called behaviors, values, or attributes. The Task Force chose to call them all skills to simplify the communication.

## Assessing Employability Skills

TABLE 3a.  
Seven Most Critical Skills  
(mean ratings less than 1.5)

ITEM NO.	ITEM	MEAN	AREA
66.	Be free from substance abuse	1.2	Personal Management
57.	Demonstrate honesty and integrity	1.2	Personal Management
1.	Pay attention to the person speaking	1.4	Academic
3.	Follow directions given verbally	1.4	Academic
54.	Show respect for others	1.4	Personal Management
55.	Show pride in one's work	1.4	Personal Management
58.	Punctual and in attendance	1.4	Personal Management

Table 3b continues the dominant trend of Personal Management Skills (50% of the table). The three Teamwork Skills (dealing with cooperation and representing the organization) combined with the Personal Management Skills account for 67% of this table. The academic skills here, just as in Table 3a appear to have a Personal Management aspect to them as well.

TABLE 3b.  
Twelve Next-Most Critical Skills

ITEM NO.	ITEM	MEAN	AREA
2.	Ask questions to clarify understanding	1.5	Academic
56.	Show enthusiasm for work	1.5	Personal Management
67.	Work productively w. min. supervision	1.5	Personal Management
4.	Answer questions accurately	1.6	Academic
9.	Read simple instructions	1.6	Academic
60.	Show initiative	1.6	Personal Management
61.	Meet or exceed work requirements	1.6	Personal Management
62.	Complete tasks under stress	1.6	Personal Management
64.	Demonstrate self-control	1.6	Personal Management
75.	Represent organization positively	1.6	Teamwork
76.	Follow organizational rules, etc.	1.6	Teamwork
84.	Cooperate with others	1.6	Teamwork

### *The Skills Ranked as Least Critical*

The skills ranked least critical present two distinct problems; specificity, and solving immediate, short-term needs. Remember that these employers were asked to answer questions across all employees, that is, generic skills that might be considered required by all. The skills in Table 3c were all academic skills. Surely, in certain jobs such as translators, the primary skill would be knowledge of a foreign language. So we think the response scale of "need" that we used measures more specificity than what might be termed "need." Skills such as metric measurement appear in the lower ranking. From a strategic standpoint, the need for metric literacy is high and growing. It appears that employers do not perceive this as an immediate need.

TABLE 3c.  
Seven Least Critical Skills

<u>ITEM NO.</u>	<u>ITEM</u>	<u>MEAN</u>	<u>AREA</u>
25.	Calculate distance, weight, etc.	3.0	Academic
26.	Understand & apply simple prob. & stats.	3.2	Academic
34.	Apply basic knowledge of social sciences	3.2	Academic
24.	Measure using metric measuring system	3.3	Academic
33.	Apply basic knowledge of natural sciences	3.3	Academic
41.	Demonstrate computer programming	3.4	Academic
6.	Understand a foreign language	3.9	Academic

## Clusters of Skills

With so many isolated skills, a cluster analysis was performed to determine whether some of these scores could cluster into major families of scores. The cluster analysis resulted in the following 8 clusters.

Table 4.  
Average of Skills Weighted by Cluster

<u>CLUSTER</u>	<u>Mean Response</u>
C1: Communication	2.1
C2: Applied Science and Technology	3.1
C3: Quantitative/Analytical	2.5
C4: Workplace Environment	2.2
A1: Responsibility/Values	1.4
A2: Career Development	2.3
A3: Teamwork	1.8
A4: Dealing with Change	1.8

The eight areas presented in Table 4 seemed to give us the best combination of describing unique areas, yet simplifying the description of the clusters. It allowed us to analyze the survey using the clusters as the central themes. Note that our clusters were similar to the findings of Carnevale et al. (1988).<sup>5</sup>

## Analysis of Open-Ended Questions

Analysis of open-ended questions shows a different picture of skills. The open-ended responses were rated by two raters who was blind to the survey results and to the basic assumptions. They were asked to look over the responses and begin to create categories, each time building on the existing categories that they had just created. The thematic analysis of this data shows more emphasis on academic skills. Since not all respondents wrote in this section, and the respondents were coded for multiple themes, there is not a true reflection of the survey respondents. However, in discussing future occupations, the kinds of skills mentioned in Table 5 present a different picture.

<sup>5</sup>Carnevale, A.P., Gainer, L.J., & Meltzer, A.S. (1988). Workplace Basics: The Skills Employers Want (Publication No. 0 - 225-795 QL.2). Washington, DC: U.S. Government Printing Office.



TABLE 5  
Regrouping of Thematic Analysis of Open-ended  
Question

<u>Theme</u>		<u>Number of Responses</u>
Academic		
1.	Spoken Language	511
2.	Read	562
3.	Write	241
4.	Basic Arithmetic	364
5.	Specialized Knowledge	83
6.	Use Tools & Equipment	505
7.	Solve Problems	240
Total Academic:		2,506
Personal Management-General		51
8.	Job Related Interests	61
9.	Values & Ethics	524
10.	Responsible	382
11.	Learn New Skills	681
Total Personal Management:		1,699
Teamwork-General		314
12.	Goals etc. of Group	97
13.	Work Teams	87
Total Teamwork:		498

### *Limits of this Study*

The response rate of 37% leaves open the doubt of how representative the sample is. We did not consider a sample of 2700 respondents to be a small number of respondents. An analysis of pilot studies of the survey did not show any bias but non-response bias cannot be ruled out.

After analyzing the results, we interpreted the responses to reflect more specificity than "criticalness" of the need for the skill. We feel further surveys that address both the need, and the prevalence of the skill are needed. We also noted that there was more variation in academic scores as if there was more agreement on Personal Management and Teamwork, and that Academic Skills may just naturally vary in need. A similar study that addresses more specific occupations (instead of filling out the survey across all occupations) would probably yield more reliable and valid results. It also seemed to us that respondents were very concerned about immediate needs instead of future needs. Maybe this accurately reflects their thinking. But, there was an indication in the open-ended responses that asking them to think about future skill needs presents a very different set of responses. The last problem we have with surveys of this type is the demographics of the respondents. The respondents in this survey were heads of personnel departments, owners (in the case of small businesses) and CEO's in the smaller businesses that had no personnel departments. It is unclear how much these respondents could adequately estimate the skill needs across an entire organization.

## Assessing Employability Skills

In summary, we believe that the insistence of the Task Force on focusing on the three areas of Academic, Personal Management, and Teamwork seems justified by the responses of this survey. We feel we can tell the public with some confidence, the skills in the survey are many of the skills needed by Michigan employers. The results of the survey and continuing work on refining our understanding of the skills will serve as a guideline for assessment and training development.

### Implications for Assessment

We are trying to organize our assessments around the three major skill areas of Academic, Personal Management, and Teamwork Skills. We are most comfortable with measuring the academic skills, and there are many tests that adequately measure these. There was no need seen to reinvent assessment in these areas. The only modification we are considering is some change of existing state educational testing to review some aspects of the content to see if they can be made more suitable to the concept of employability. The Personal Management and Teamwork skills were more problematic. Our initial review of the literature<sup>6</sup> found that the reliability and validity of instruments were not as good as academic measures. Our initial view of these areas were that we would look for assessment that we could rate personal management skills behaviors and we could observe teamwork behaviors.

Because of the complex nature of the assessment requirements (in terms of the three dimensions, and the depth of skills within any given area and the variety of different skills needed, depending on the career aspirations of the client), we have focused our efforts on building a prototype for some type of record keeping system or "portfolio" that would allow multiple assessments in each of the three areas. The state is now beginning to work with referent groups to define standards for what that portfolio might look like, and how the portfolio could be used to equate job skills with career goals and training needs. We are also surveying the best assessment practices used in this state and used nationally that focus on these three areas. We are hoping that the portfolio will be individualized and provide for no ceiling of skill levels and no minimal levels. It should reflect the individuals actual skill levels and serve as an organizational tool to use in career planing and reviews to build a resume and job skill inventory for the employer.

---

<sup>6</sup>Miller, J.V., & Pfister, L.A. (1988). *Employability Skills Assessment Literature Review*. Prepared for the Michigan Department of Education, Office of Technical Assistance & Evaluation, Michigan Educational Assessment Program.

### A Unique Solution

#### *The Michigan Opportunity Card*

The paperwork needed to monitor state programs and individual information for the wide variety of programs, community colleges, other state and private training efforts, and employee skills presents a formidable barrier to effective coordination and progress. To meet this challenge, the Governor's Commission on Jobs and Economic Development has proposed the introduction of the Michigan Opportunity Card. This is a credit-card-size "smart card" with a computer chip to store thousands of pieces of information. The card may be inserted into a card-reader which will tell everything that is needed for the involved agency to determine the individual's needs, strengths, weaknesses, financial standing, etc., relating to the program (s) of that agency.

Soon, it is planned to have M.O.C.'s offered to all-graduating seniors, Michigan youth corps enrollees, and dislocated workers. Core services would include:

- a. inventory of job training, education, and employment related community support services;
- b. assessment of cardholder's basic academic and occupational skill needs, interests, employability, and achievements using valid, reliable, standardized and appropriate instruments;
- c. eligibility for funding and/or services;
- d. individualized personal plan of action (PPA) for achieving and managing employment, training, and education goals;
- e. basic employability skills training to help every adult get the academic, personal management and teamwork skills necessary for existing and future jobs;
- f. job placement assistance in the future, the results of a prospective employee's employability skills assessment may also be stored on the card. Prospective employers would be permitted access, on a "need to know" basis, to relevant, public information as released by the candidate. Full consideration of privacy requirements would be attended to in the access codes of the employer and employee, to protect the individual's rights.

The M.O.C. could be integrated with the portfolio concept with card being the electronic form. Refinements in both the card and the experience with the portfolio will compliment one another.

### Conclusion

New assessment techniques need to be addressed. The curriculum in the schools and in job training programs must be looked at in terms of the employability skills profile. It appears that we have spent a great deal of time using assessment methods for the selection of employees and as the demographic patterns in the United States change, we see an increasing need to link assessment practices more closely with development of skills in the individual.

# JOB ANALYSIS, PERFORMANCE APPRAISAL, AND ORGANIZATIONAL JUSTICE

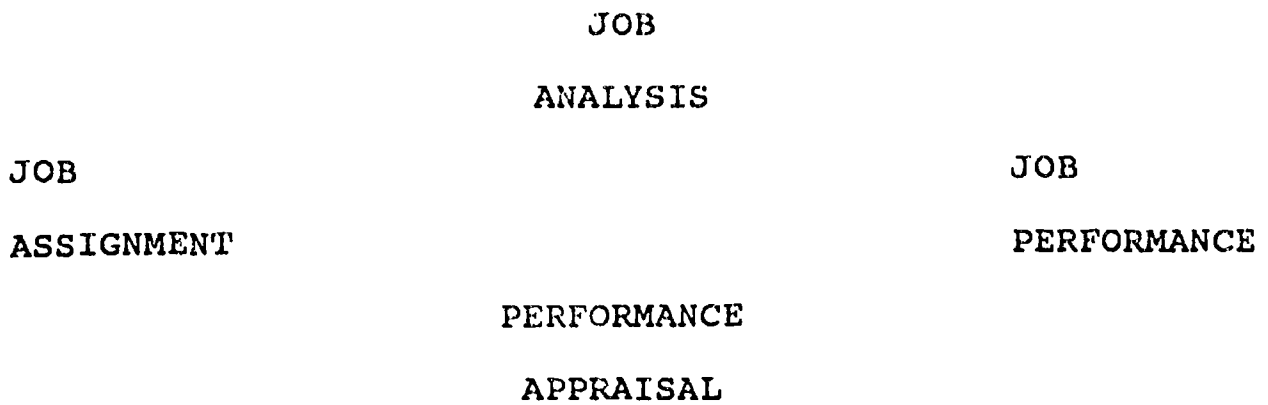
BY

THOMAS J. ATCHISON

## INTRODUCTION

Nobody seems to like their performance appraisal system. Employees complain that the results are unfair and don't represent their true performance. Supervisors feel that it takes up too much of their time and doesn't leave them with usable results. Human Resource professionals feel that the supervisors don't pay enough attention to doing it right and the results aren't what the organization needs to make the personnel decisions that rest on performance appraisal.<sup>1</sup> Why does such a natural part of the managerial job go so wrong when the organization tries to systematize the process? This paper explores some aspects of this question.

Job analysis and performance appraisal are considered to be the heart of Human Resource Management. Information obtained from these two processes is used in all other human resource activities. In particular, these two processes are a part of the cycle of work by which human effort is translated into the accomplishment of organizational goals. Figure 1 is an illustration of this cycle. The cycle starts with the assignment of work [job design]. This assignment is examined and described by job analysis so that both the organization and the person are clear as to the assignment. The employee then performs the work and the supervisor appraises the results of the work. This latter information is used as feedback both to the employee, as to how well s/he did, and the supervisor



## A WORK CYCLE MODEL

FIGURE 1

in assigning future tasks.

This cycle of work is a major organizational process. A process in this context being a sequence of interrelated events to accomplish some organizational purpose.<sup>2</sup> Job analysis and performance appraisal are sub-processes which support this larger process. Organizational processes use systems to accomplish their ends. Systems in this sense are ways of doing things, such as policies and procedures that are intended to aid and control the processes in accomplishing the goal.<sup>3</sup> But all these processes and systems need to work smoothly together if the total processes is to operate correctly. In the case of the work cycle there are a number of ways in which the interface between job analysis and performance appraisal does not integrate properly. These problems will be discussed in two directions, first from job analysis to performance appraisal and then in the opposite direction.

### FROM JOB ANALYSIS TO PERFORMANCE APPRAISAL

Job analysis is a process by which information about jobs is collected and recorded. It is a descriptive process that precedes

the action to be taken but is non-evaluative in and of itself.<sup>4</sup> The results of job analysis however are used to evaluate both jobs and employees. The focus in job analysis is on collecting the proper information to make these further judgments. In order for job analysis to be of use to performance appraisal information such information must be collected and recorded. The deficiencies in this area are the topic of this section.

**The Absence of Performance Standards** The end product of job analysis is a job description. This description is used for many Human Resource Management purposes but a major purpose is to describe to the employee what s/he is to do and what the performance standards are by which task accomplishment will be judged. Clearly then the job analysis process should collect performance information, not about how the employee has done but about how the employee is expected to do. In this regard job analysis as practiced does not accomplish its goal. Typically job descriptions do not contain information about performance standards. Why? First, most job analysts see performance appraisal as a separate process that has little to do with the description of the job. Second, defining exact performance standards for tasks is a hard job that neither the job analyst nor the supervisor wish to spend a lot of time on.

While collecting and defining performance information is difficult, the lack of such information in job descriptions may be viewed as a violation of organizational justice. The job description is the major source of information for the employee as his/her obligations under the employment contract. To state job



descriptions only in terms of what is to be done and not in how those activities are to be judged is unfair if, as already stated, performance appraisal is used to make many decisions affecting the status of the employee in the organization.

What needs to be done? Performance standards can be most clearly stated when job descriptions are based upon tasks performed. If each task is described in terms of its "what, why, and how" the basis of establishing performance standards is available. The "why" part of the task statement can be used as a basis for thinking about the appropriate performance standards. For example, take a task statement that says:

"Lectures students to dispense information".

The purpose of this task, from the organization's perspective, is "to dispense information" and the statement can now be used to define a performance standard consisting of a measure of the amount and quality of information dispensed in an appropriate time period. This may not be an easy task but it is well worth the time and effort required.

**Ambiguity of Performance Standards** The lack of performance information in job descriptions creates an ambiguity as to what exactly are performance standards. Conceptually, three things could be measured that is included as performance, the outcomes, the behavior, or the inputs.

Ideally all jobs could state some performance standards in terms of the intended outcomes of the job. Some jobs do have clearly measurable performance outcomes, such as sales persons or assembly workers. But many jobs do not have such clear cut outcomes

nor are the outcomes independently arrived at. For instance, in the example given above of lecturing an appropriate outcome measure would be that of student learning. The problem is that while that is what the organization wants the lecturer has little control over that outcome. A focus upon outcomes would dictate that performance appraisal should take on the characteristics of management by objectives [MBO]. But, if performance standards are tied directly to the tasks to be performed as suggested above then the MBO has a more enduring basis than just the bargaining between supervisor and subordinate.

It may not be enough to measure just outcomes from the job nor may it be possible. The employee may accomplish the right things in the wrong way. Thus, the behaviors of the job are also important. In terms of the example above how the instructor lectures may also be important. The way tasks are accomplished coordinate the employee's goals and the organization's goals more closely since individual jobs are not done in a vacuum but in concert with other jobs and in an environment which demands that things be done in certain ways. From the viewpoint of organizational justice, behavior criteria need to be included to prevent the focus on outcomes forcing the employee into actions that are illegal or immoral.

The third possibility is that performance standards have to do with employee inputs. This alternative is the least appealing for the purposes of the work cycle discussed in this paper. Such an approach suggests that the employee inputs will lead to particular behaviors, which in turn will accomplish specified

outcomes; a tenuous sequence at best. This approach leads to the least desirable aspects of performance appraisal, that of the supervisor judging the value of the employee's personal characteristics. If a goal of performance appraisal is change then this approach is also contrary to organizational justice in the sense that it is intrusive upon the employee, as a person. finally, the use of input factors is illegal when the performance appraisal is used to validate a selection system.<sup>5</sup>

**Discretion in job assignments** A final aspect of looking from job analysis to performance appraisal has to do with the nature of the job assignment. Performance appraisal assumes that the employee's efforts can make a difference in how well the task is accomplished. There must be discretion in how the tasks are carried out. If this is not true it is unfair to hold the employee responsible for the outcomes. The responsibility rests entirely with the supervisor or those who designed the jobs for the outcomes which occur.<sup>6</sup>

Operationally this gets translated into how we describe the tasks of the job. Often jobs are described not in terms of tasks but in terms of responsibilities which are stated in very general terms. Thus, in our example above the instructor may have a statement that s/he is responsible for student learning. Such is not the case as already discussed. The instructor is responsible for certain things which are intended to aid student learning.<sup>7</sup>

#### **FROM PERFORMANCE APPRAISAL TO JOB ANALYSIS**

Not only does job analysis ignore performance standards but the development of performance appraisal systems in turn ignore the

nature of the jobs in the organization. This section explores some aspects of not focusing on the job when developing performance appraisal systems.

**The Conflict of Goals** In performance appraisal there is a schism between the process and the system. The systems that are intended to support the performance appraisal process instead take on a life of their own intended to achieve different goals. Part of this confusion comes from the different goals that performance appraisal is intended to accomplish. One goal is associated with the work cycle, the accomplishment and improvement of work. A second goal is administrative. Organizations use performance appraisal data to make decisions about employees and to judge the value of other Human Resource systems, such as staffing.<sup>8</sup>

The requirements of these two goals are not only different they are conflicting. The work cycle goal is clearly related to the job to be done and the judgments are idiosyncratic to the particular job and person. The administrative goal requires a standardization of systems so that comparisons and measurements can be made. The result is that performance appraisal systems are devised to be organizational measures of performance, not job measures of performance. The most common form of performance appraisal system is a graphic rating scale on which all employees are rated using a common set of factors. These factors are only partially job related and are not arrived at on the basis of job analysis.

Using this system the employee and not the job becomes the focus of performance appraisal. The concept of performance

standards by which the employees is judged is replaced with an examination of the employee and his/her characteristics. This is uncomfortable for and unfair to both the employee and the supervisor. The employee feels that s/he is being judged on the basis of personality and not in terms of what s/he has accomplished. After all a poor match between personal characteristics and the job is a matter for staffing and not performance appraisal. As for the supervisor, s/he is asked to make judgments about the person and not the behavior or outcomes of the job assignment. Supervisor often express this frustration as being asked to "play god."<sup>9</sup>

**Performance Factors are Non-Job Related** The adaptation of the actual performance to non-job related performance factors leads to an imperfect relationship between performance and its appraisal. The supervisor must make a "leap of faith" in relating the factors on the performance evaluation form with the performance as s/he has observed it. From the employee's standpoint it appears that s/he is being rated on hidden factors since there is this unspoken translation that is taking place. This difference in job based performance and the performance appraisal form is another source of injustice.

The use of standardized performance appraisal systems assumes that there is only one form of good performance. In fact, "doing a good job" means different things for different employees. Any supervisor has employees who are good and not so good at different aspects of their jobs. It is the supervisor's and employee's responsibility to work out these strengths and weaknesses in such



a way that the strengths are used and the weaknesses minimized. Having done this it would be unfair to come along at performance rating time and rate the employee low on a factor that both had decided had been worked around. In this regard the halo effect may be a positive aspect of performance appraisal, not a negative one.

**Supervisory Skill and Motivation** Supervisors are under a lot of conflicting pressures in performance appraisal. It is assumed that they have both the ability and motivation to do this aspect of their job well, but they don't.<sup>10</sup> Having the ability to do a good job of performance appraisal assumes, among other things, that the supervisor has all the information needed to make a proper judgment and knows how to use that information. In fact, supervisors often do not have the necessary information and are embarrassed to admit that they do not have it since the perception in organizations is that supervisors should know everything about what their subordinates are doing. Given what was said above it should also be clear that the job of translating the observations made about an employee's performance into the format of the organization may take considerable training.

Supervisors may not have the motivation to do performance appraisal well, or at least what the organization considers to be performance appraisal. When the organization's performance appraisal system has little to do with the everyday performance appraisal process of the supervisor why would s/he be motivated? One way of expressing this is that the supervisor feels that the system "belongs" to somebody else, usually the Human Resource Department. Performance appraisal is a lot of work for which there

is little or no payoff [or perhaps negative payoff].<sup>11</sup>

**Performance Appraisal as Control** In the work cycle model presented at the beginning of this paper performance appraisal feeds back to the beginning of the cycle and determines the next assignment of work. In this respect performance appraisal is like a control mechanism in the organization and an observation that what happened is not what was expected or desired is used to make corrections in behavior. Thus, the emphasis in performance appraisal is upon feedback and improvement of the employee's performance. But this is not what sometimes does or should happen.

In control systems, deviations from standard may mean a change in behavior is required or it may mean a change in the standard is required. Similarly, in performance appraisal a difference between expectations and actual may call for a change in the performance standard, not in the behavior of the employee. This kind of response is not considered in performance appraisal systems. In staffing organizations the emphasis is upon matching the person to the job. The job is considered the constant and the person the variable. It may be more practical when dealing with employees already selected into the organization to think in terms of matching the job to the person. This places an emphasis in performance appraisal of examining each employee's strengths and weaknesses and working out how to best design work within these parameters. Performance appraisal training programs with this as a focus would be far more productive than ones attempting to reduce rater bias.

Providing feedback to employees is supposed to provide them with the information and motivation to change their behavior. But it may not have that response at all. In fact what is likely to happen is a downward spiral of the work cycle. If performance is seen as below standard by the supervisor two things are likely to happen. One, the supervisor is likely to lower the level of the work assignment in the next cycle. Two, the employee rather than changing is likely to accept that evaluation of him/herself and in consequence not try to change but adapt to the lowered performance requirement. Thus, performance appraisal becomes a self fulfilling prophesy. Those who do well see themselves as doing well and strive harder while those that don't do well accept that they are not as good. Where the focus in performance appraisal is on personal characteristics rather than the results of the work this downward spiral is exacerbated. This can be a serious organizational justice problem when the basis for the judgments about performance were clouded by the sex or race of the employee.<sup>12</sup>

**The Time Frame of Performance Appraisal** Viewing performance appraisal as a part of an ongoing work cycle puts a different focus on it. Performance appraisal systems focus on discrete past time periods, usually six months. Thus, the performance appraised is just the last six months, a "what have you done for me lately" approach. But the employment contract does not run in discrete six month segments. It is an open ended contract that contains features of the past, present, and future. So performance needs to be looked at more broadly as including past contribution [which employees may feel have not been fully reimbursed], present performance levels,

and future promise. This approach joins performance appraisal and career advising and takes the focus from just the present job to the relationship of the employee to the organization, not just with a particular supervisor. In fact, most supervisors are not well trained to provide this kind of performance appraisal.

## CONCLUSIONS

This paper has focussed on the relationship between job analysis and performance appraisal in the context of organizational justice. What comes out of this comparison is the following:

1. Performance appraisal is a very difficult process for supervisors and employees bc . in terms of doing it and accepting the results of it. The systems used to accomplish it often sub-optimize the goals of the process.

2. The relationship between job analysis and performance appraisal is tenuous at best. Neither fully uses nor supports the other. This leads to performance appraisal systems which do not do what they are designed to do and are not grounded in the work of the organization.

3. Job analysis and performance appraisal are both sub-processes of the work cycle in the organization. This relationship is not clear and leads to systems in both areas which do not support the other nor help the supervisor manage the work of his/her unit.

4. Performance appraisal needs to become more focussed on the individual job and its relationship to the organization's goals and less on the employee's characteristics.

1. Geis, A.A. "Making Merit Pay Work" Personnel, Jan. 1987, pp. 52-60 and Lefton, R.E. "Performance Appraisal: Why They Go Wrong and How to do Them Gight" National Productivity Review, 1986, 5(1), pp. 55-63.
2. French, W. L. Human Resource Management, 2nd. Ed., Boston, Houghton Mifflin Co. 1990, p. 10.
3. Ibid.
4. Ghorpade, J. and T.J. Atchison, "The Cocept of Job Analysis: A Review and Some Suggestions" Public Personnel Management, 1980, Pp. 134-144.
5. Cascio, W. and H.J. Bernardin, "Implications of Performance Appraisal Litigation for Personnel Decisions" Personnel Psychology, 1981, pp. 211-26.
6. Brown, W. "What is Work" in Brown W. and E. Jaques Glacier Project Papers, London, Heinemann Books, 1965, Pp. 54-73.
7. Jaques, E. Equitable Payment, Carbondale, Southern Illinois Press, 1970.
8. Meyer, H.H. E. Kay and J.R.P. French, "Split Roles in Performance Appraisal" Harvard Business Review, 1965, Pp. 123-129.
9. Rice, B. "Performance Review: The Job Nobody Likes" in Ferris, G.R. and K.M. Rowland, Human Resource Management: Perspectives and Issues, Boston, Allyn and Bacon, pp. 163-69.
10. Heneman III, H.G., D.P. Schwab, J.A. Fossum, and L.D. Dyer, Personnel/Human Resource Management 4th Ed., Homewood, Ill., Irwin, 1990, pp.160-3.
11. Rice Op Cit.
12. Kord, J.K., K. Kariger, and S.L. Achechtman, "Study of Race Effects in Objective Indexs and Subjective Evaluations of Performance: A Meta Analysis of Performance Criteria" Psychological Bulletin, 1986, Pp. 330-7.



## Test Item Design and Evaluation

Dr. Thomas M. Haladyna  
Arizona State University West  
PO Box 37100  
Phoenix Arizona 85069-7100

### ABSTRACT

The concepts, principles, and procedures of designing objectively-scorable test items is not yet a science, but recent progress is encouraging that item writing can be more than a distillation of wisdom passed on from mentors, textbooks, and experience. On the other hand, the evaluation of test items is becoming a science, and there are numerous theories and methods for studying, evaluating, and improving test item. This paper reviews recent progress, provides recommendations, and indicates work that is needed to advance the sciences of item design and evaluation.

Objectively scorable testing continues to be the main way test users measure achievement and ability (aptitude). Achievement is viewed as cognitive behavior which is modifiable through experiences, more specifically instruction; ability is viewed as cognitive behavior which is resistant to modification through experience but is useful in predicting mental capacity or potential to learn. The main reason for using objectively-scorable test formats is that it is the single most effective way to measure knowledge.

Despite increases in emphasis on performance testing, there are many areas in which objectively-scorable testing continues to be useful. Some of these are: (1) personnel selection, (2) instructionally-based achievement testing, (3) licensure and certification examination, (4) admissions testing, and (5) research and evaluation.

Critics of multiple-choice testing have decried the woeful scientific basis for the design of test items (Cronbach, 1970; Haladyna and Downing, 1988a; Nitko, 1984; Roid and Haladyna, 1982). Most techniques of item writing are based on folklore consisting of experience or wisdom passed on by teachers or mentors. Fortunately, there have been some advances to report, but not to the extent to which that multiple-choice test item writing can be accepted as a science. On the other hand, item analysis and evaluation (working in the theoretical frameworks of classical theory, item response theory, generalizability theory) has become a highly sophisticated endeavor.

This paper focuses on two important activities in test development: the design and evaluation of test items. The emphasis will not be on traditional practices but on recent innovations both in thinking and in practice.

### Advances In Item Writing

#### Theories of Item Writing

Roid and Haladyna (1982) reviewed and evaluated a variety of theoretical approaches to item writing. These theories are enlightening and important, because each represents an attempt to make item-writing a science worthy of the many sophisticated item analysis techniques that have been created in the past decade. These theories are useful in many ways. First, they emphasize the operational definition of knowledge to be tested. This formalization of content is often viewed in terms of content domains, which actually define the types and extent of cognitive behavior upon which tests and test items are based. Second, the use of these theories removes the idiosyncrasies of item writers by developing rules that item writers must follow. Third, a comprehensive set of items is generated which, by the nature of the theory and item-generating rules, specifies the complete domain being measured. Fourth, the item-writing procedures provide evidence of construct or content validity by virtue of its item development procedures and the theory upon which items are based.

Most noteworthy among these theories was the algorithmic approach of Bormuth (1970). He developed an 82-step algorithm for translating prose into objectively scorable test items. His theory was extensively tested by Roid and his colleagues (see Roid and Haladyna, 1982). The obvious advantage of advancing his theory is to test comprehension from textbook instruction in a relatively automated manner. Computer-driven item generation is entirely possible by programming a modified version of this algorithm for the computer so that text can be read into the computer and test items can be produced.

Guttman (1969) suggested an automated item-writing method based on facet theory. Briefly, facet theory allows the definition of an ability or achievement domain through the use of mapping sentences. Item-writing can be regulated by specifying the content domain to be tested via statements that contain facets. A simple mapping sentence containing a single facet would be:

The (capital/largest) city of any of 50 states is ...

There are 50 capitals and 50 largest cities. In some circumstances, the capital is also the largest city (e.g. Phoenix, Arizona). Thus the range of possibilities for a single set of test questions based on this statement is less than 100.

Hively (1974) introduced a method similar to facet theory. It too offers item forms, similar to Guttman's mapping sentences, but seems best suited to quantitative and technical content. An example of an item and a derived item from are taken from Roid and Haladyna (1982, p. 118):

A random sample of 100 trucks weighed at a highway checkpoint had an average of 40,250 lbs. with a standard deviation of 2,500 lbs. Find the 95% confidence interval for the true average gross weight of trucks passing this checkpoint accurate to at least one decimal place.

Item Form:

Given a random sample of (N) (objects) with an average (dimension or feature) of (M), with a standard deviation of (SD). Find a (95%, 99%) confidence interval for the true average (dimension or feature) of the (objects) accurate to at least one decimal place.

The item form merely generalizes the above item into a myriad of possible alternate items which facilitates the development of a test or sub-test over this quantitative problem-solving skill.

The quest for how to measure higher level thinking has led Tiemann and Markle (1978) to a system for defining concept attainment and a measurement process as well. Their system lends itself nicely to cognitive learning theories and the way the mind processes and uses information in thinking.

Williams and Haladyna (in Roid and Haladyna, 1982) present LOGiQ (Logical Operations for Generating Intended Questions), a system which provides a classification matrix as well as generic objectives that fit the system. Thus items can be generated from objectives once content has been identified to match these objectives. One advantage of this system is the variety of higher level thinking categories that is crossed with content categories.

While these theories are a step in the right direction for item writing, none of these theories have advanced, and the labricious nature of most of these theories' technologies appears to limit their eventual usefulness. Still the existence of these theories offer hope that item-writing may someday become a science, and that item writing will be free of the flaws and faults that seem to limit test development today.

### Innovative Formats

The traditional multiple-choice format has existed for quite some time. Generally, it has a stem, usually stated in a question format or a partial-sentence format. Four or five options are generally recommended by textbook authors (Haladyna and Downing, 1988a), but there seems to be ample evidence to suggest that three options are sufficient for most testing situations (Haladyna and Downing, 1988b).

Quite naturally, test and item developers have sought better ways to test knowledge. The matching format (see Table 1) is a slight modification of the multiple-choice where a single set of options precedes the stems. This modification provides focused, concentrated testing of a single set of concepts in a highly efficient manner. The true-false format has also existed for a long time, but has a negative reputation, largely due to the misapplication of it to measure factual recall. For many years, Ebel (1979) championed the true-false format, illustrating that indeed it could be used successfully. Unfortunately, a recent review by Downing (1990) offers evidence that true-false format is typically inferior to more conventional multiple-choice testing.

There are, however, a variety of proposed formats which are multiple-choice. Each represents a unique way to test knowledge, and each has a small body of research which recommends its use, except for the complex multiple-choice (Type K), which is not recommended.

Alternate-choice. The alternate-choice format is two-option multiple-choice. One merely writes an item stem and provides a right answer and a plausible, grammatically parallel wrong answer. Table 1 provides an example. The major limitation of this option is the tendency for someone not knowing the right answer to guess the right answer. However, this tendency is more than offset with the efficiency of the format. One can ask a multitude of alternate choice items in a set time period, say one hour. Further, the floor of the test scale based on two options must be 50%, so interpretations can be made with this in mind.

There is a strong rationale for alternate-choice testing. Haladyna and Downing (1988) examined a variety of professional tests and evaluated distractors to four- and five-option items. They found that while these tests were of high quality and adhered to high psychometric standards, the average number of good distractors per item ranged between one and two. Thus, if one discarded useless distractors, as judged by criteria used in their study, most test items would contain two- or three-options per item.

There is theoretical support for using two-option items. In the context of item response theory, Lord (1977) showed that two-option testing is ideal for high achieving examinees, while four- or five-option testing may work best with low achieving examinees. This conclusion is credible when examining the results of the Haladyna and Downing (1988) study, because with high achieving students, they seem to reduce choices to one or two options, while with low achieving students, they cannot reduce choices to two options. Finally, in a recent review of the alternate-choice format, Downing (1990) concluded that this format is useful; that research modestly supports its use. The ease of writing alternate-choice items is unquestioned.

Complex-multiple-choice. As shown in Table 1, the complex multiple-choice, sometimes referred to as Type K, has a set of options which are reconfigured into combinations which become the actual multiple-choice options. Criticism has been mounting against the use of this format (Albanese, 1990; Haladyna and Downing 1988a; 1988b), and at least one testing agency, the National Board of Medical Examiners, has discontinued the use of this format. Therefore, it is recommended that this format not be used in the future because these kinds of items are usually more difficult, less discriminating, and require more development and administration time.

Multiple true-false. This format, also shown in Table 1, is a marriage of true-false and multiple-choice formats. The item takes the form of a multiple-choice format, but the examinee is expected to evaluate the truthfulness of each option. Thus a conventional four-option multiple-choice item can be easily converted to a multiple true-false item by asking the examinee to mark A if true or B if false for each option. Notice that each option is numbered because each is a true-false item, while the stem is not numbered because it is merely the stimulus for the items. Frisbie (1990) presented positive research findings on this format and endorsed the use of this format. However, he also points out obstacles to its use. The greatest of these is the test user's unfamiliarity with the format. Another problem may be local dependence, the tendency for one item to influence the response to another. There is little doubt as to the efficiency of this format. Finally, any complex multiple-choice item can be easily modified into a multiple true-false item.

**Table 1**  
**Examples of Multiple-Choice Item Formats**

**Conventional Multiple-Choice**

For what is San Diego best known?

- A. Outstanding restaurants and fine dining
- B. Mild climate and recreational opportunities
- C. Major league baseball and football teams

**Matching Format**

Match the city with a distinguishing feature.

- |  |  |
|--|--|
| <ul style="list-style-type: none"> <li>A. San Diego</li> <li>B. San Francisco</li> <li>C. Seattle</li> <li>D. Los Angeles</li> </ul> | <ul style="list-style-type: none"> <li>1. Proximity to islands and protected ocean water</li> <li>2. Most variety of entertainment and tourist attractions</li> <li>3. Cool climate and unpolluted air</li> <li>4. Popular for ease of transportation and many tourist attractions</li> <li>5. Considered the most livable for these cities</li> </ul> |
|--|--|

**Alternate-Choice**

What is the primary attraction in downtown San Diego?

- A. Horton Plaza
- B. Gaslight District

**Complex Multiple-Choice**

Which of the following best represents San Diego's attractiveness?

- 1. Climate
  - 2. Beaches and water sports
  - 3. Tourist attractions (e.g. Sea World)
- 
- A. 1 & 2
  - B. 2 & 3
  - C. 1 & 3
  - D. 1, 2, & 3

Table 1 (continued)

**Multiple True-False**

Which of the following are major attractions in San Diego?

1. Sailing and water sports
2. Restaurants
3. Shopping
4. Tourists attractions (e.g. Sea World)
5. San Diego Zoo
6. Climate
7. Low hotel rates in summer
8. Low cost of living
9. Excellent public transportation

**Item Set (Testlet)**

You are planning a one-week vacation to a West Coast city. Your main interests are reasonable cost of living, mild summer climate, sailing, sightseeing, mountain hiking, and shopping.

1. What is a good rule-of-thumb regarding daily lodging and food costs for a party of two?
  - A. \$50/day
  - B. \$100/day
  - C. \$200/day
2. If renting a sub-compact car for a minimum cost, what is a reasonable guideline regarding cost?
  - A. Weekly rate of \$120
  - B. Weekly rate of \$150
  - C. Weekly rate of \$175
3. Which city would you choose?
  - A. Seattle
  - B. San Diego
  - C. Los Angeles
  - D. San Francisco



Context-dependent item set. The motivation to measure higher level thinking has in part led to the development of this item format, which, like others, is shown in Table 1. The context-dependent item set consists of stimulus material and a set of 5-12 related test items. The test items may be any format previously discussed and additional any open-ended question format (i.e. short-answer essay, long-answer essay, completion).

Wesman (1971) identified three types of context-dependent item sets, each is distinguished by the stimulus material that precedes the items: (1) pictorial form (which includes pictures, maps, drawings, graphs, data, photographs, works of art, etc.), (2) interlinear, which consists of a single written passage and a number of denotations which provide a basis for questioning, usually applied to the detection of grammatical, spelling, punctuation, and capitalization errors, and (3) interpretive, which is used to test reading comprehension. Haladyna (1990) adds a fourth type, the problem-solving exercise which presents a problem in the stimulus material and then provides a battery of questions aimed at testing various aspects of problem solving. Research has been very limited in this area, but the format is very popular in some certification and licensing testing programs (Haladyna, 1990).

Haladyna (1989) introduced a method for developing context-dependent item sets for the teaching of beginning and intermediate statistics. This procedure is loosely based on Guttman's facet theory (discussed earlier in this paper), but is very efficient in generating test items due to the generic structure of the problem-solving skill required. The main feature of this approach is that inefficiency in writing the scenario and related test questions is eliminated by making the development of statistical scenarios generic in nature. Hundreds of items can be generated by making slight changes in base scenarios.

While the format is inefficient to construct and administer, it appears to be very useful for measuring complex cognitive behavior. It is expected that this format will be expanded in variety and increased in usage because of the urgent need to objectively test for problem solving skills and other types of higher level thinking (Nickerson, 1989). One major limitation of this technique is the general lack of existence of a prevailing paradigm for higher level thinking. The Bloom taxonomy has existed since the mid 1950s, but scientific evidence has never really been convincing that the taxonomy accurately reflects human cognitive behavior (Seddon, 1976).

### Item Shells

An item shell is a hollow item, that is, an item that has had its content removed (Haladyna and Shindoll, 1989). The item shell is the syntactic structure from a successfully performing item that measures a generalizable cognitive behavior from a test domain. The item shell technique is loosely based on the work of Guttman (1969).

The development of an item shell is quite simple, which makes the technique appealing to content experts who lack experience in item writing. The method's simplicity also helps alleviate "writer's block" which is the malady that inhibits item writers from writing the item. The item writer identifies a successfully performing item that measures a generalized cognitive behavior of importance in the test. For example, in a high school consumer mathematics class, the teacher may want to test for mathematical problem solving of a practical nature involving financing the purchase of an automobile. The original question states:

What is the annual interest charge on an auto loan of \$10,000 at 8.5%?

If the question is a good performer, and the learning objective requires the calculation of interest, we can strip out the loan amount and percentage rate and replace it. By varying these two variables, a large number of similar items can be generated.

If an item tests understanding of a concept:

What is the distinguishing characteristic of a bird?

By removing the word "bird," we can generate additional item by substituting such terms as "mammals," "amphibians," "reptiles," and "fish."

Haladyna and Shindoll (1989) provide a variety of item shells for medical problem solving that tap different aspects of higher level thinking, and the approach is useful for any content area where multiple-choice items are used. However, the item shell applies equally well to open ended test items, such as completion and short-answer essay. Table 2 provides a list of generic item shells, all of which were derived from successfully performing test items in various certification and licensing testing programs.

### The Role of Cognitive Learning Theories In Item Writing

Considerable interest has been mounted in the specification and measurement of higher level thinking in education, both for achievement and ability testing (Nickerson, 1989). Another influence is the strong interest in merging cognitive psychology with testing (Snow and Lohman, 1988). Public education is clearly allocating considerable resources to help its students become thinkers and problems solvers in the 1990s instead of memorizers. In the business sector, companies like Motorola Corporation are considering ways to train employees to problem solve on the job rather than to perform automated mental functions that are essentially low-level thinking. Businesses seem to need workers who can identify problems and solve problems rather than to simply perform low level activities repetitiously. The demand is considerable for objectively scorable methods for measuring higher level behavior in achievement or ability. The continued study and use of these innovative item formats and the development of theories of item writing are necessary. However, the isolation of research and development in cognitive psychology, as with the isolation of research

**Table 2**  
**Generic Item Shells**

**Defining**

What is the meaning of...  
What is synonymous with...  
What is like...  
What is typical of...  
What is characteristic of...  
What is an example of...  
What distinguishes...  
What is the definition of...  
What is a good example of...

**Predicting**

What would happen if...  
When..., then what happens?  
If..., then what happens?  
What is the outcome of ...?  
Under these circumstances..., what would you expect?  
What are the expected findings for...

**Evaluating**

Which is the most important...  
Which is the most significant...  
Which is the best...  
Which is the worst...  
Information is given. What is the best treatment of... in this situation?  
What is the best procedure for...  
What is the most effective...  
What is the most efficient...  
Which of the following is true?

**Applying**

What is the correct way to...  
Given information..., what action should be taken?  
In this situation..., what would you do?  
Background is given. How would you do something?  
What is the problem?  
What is the solution to the problem?  
How should the problem be solved?

and development in testing is basically unhealthy. The emergence of a unified, holistic approach to thinking about the mental processes and testing is clearly the most fruitful direction in the future.

## **Advances In Item Evaluation**

### **Traditional Item Analysis**

The most common approach to item evaluation is to compute two item characteristics: difficulty and discrimination. The difficulty is the percentage of correct responses to an item in the sample being tested. Moderate difficulty is often considered in the general range of .50 to .85. Discrimination is the correlation between item performance and total test performance. The most typical item discrimination indexes are the biserial and point-biserial correlation, two statistics which differ in conceptualization and computation but are nearly perfectly correlated to one another. Discrimination indexes that are positive, say above .20, signify the item discriminates examinees with respect to content being measured by the test. Negative discrimination for a correct answer is very bad. It can be easily shown that discriminations functionally related to test score reliability, there the maintenance of good item discrimination has a direct effect on reliability.

Item analysts generally use these two item indicators and arbitrarily conceived limits for difficulty and discrimination to identify items that are too easy, too hard, and not discriminating. Occasionally, keying errors in items can be identified via item analysis. Keying errors, if undetected, can be very detrimental to persons who test scores fall close to important decision points on the test score scale. Therefore attention should be paid to items which are suspect in this regard.

A major limitation of item analysis is that these statistics are sample dependent. If the same of examinees is unusually restricted, consisting only of high scorers or consisting only of low scorers, the resulting item statistics of difficulty or discrimination are inaccurate. Traditional item analysis, based on classical test theory (Lord and Novick, 1968) has been criticized in this regard, and item response theory (Lord, 1980), which is resistant to varying sample conditions, has been championed as its successor.

Traditional item analysis is still useful but only identifies performance problems of items in samples which vary widely in test scores. Traditional item analysis does not identify causes of problems, nor does it identify ways in which to improve items. Methods for improving or polishing items are clearly needed to advance item writing.

### **Analysis of Adherence to Item-Writing Rules**

Most standard textbooks on testing offer lists of item-writing rules and examples of well-written and poorly-written items. Haladyna and Downing (1988a) distilled 43 rules from a review of 45 textbooks. In a second study, they examined the validity of these rules from the standpoint of empirical research (Haladyna and Downing, 1988b). They found that there was indeed little support for many rules, and actually some disagreement with testing experts viewpoints, as obtained from the review of textbooks.

Nonetheless, the existence of a set of validated item-writing rules presented by Haladyna and Downing (1988a) provides some guidance to item writers regarding item writing, although the validation of many rules is still clearly needed. It also stands to reason that one way to improve test items is to perform a regular review of any new test items to determine if these 43 rules are followed. Table 3 presents a summary of these rules. Textbooks on educational measurement provide a more comprehensive treatment of this subject, but keep in mind that textbook authors are not in agreement about all item-writing rules presented in Table 3.

### **Content Review**

The advent of instructional alignment (Cohen, 1987), criterion-referenced testing (Glaser, 1963), and integrated teaching and testing (Nitko, 1988) has focused attention on the importance of linking content to tests to instruction. Related to this linkage is the need to carefully specify the content domain of any test, whether achievement or ability, and to identify items that carefully map the domain of content according to test specifications which identify both the content and the cognitive behavior. Fitzpatrick (1982) focused our attention on the importance in content validation of specifying content domains and carefully identifying items that map that domain. Content validation is a more complex enterprise than simply ensuring that a test represents a balanced sample of content from a well defined content domain. It is essential in the development of any test to ensure that several aspects of content representation are present, and this becomes part of the item evaluation process.

First, the content domain for the test must be specified in ways that the public can understand. In certification and licensing testing, more and more, this requirement has come to mean that a study of professional practice (referred to variously as a *practice analysis*, *job analysis*, *task analysis*, or *role delineation study*). In achievement testing, a content domain is defined by objectives, goals, or domain specifications. In ability testing, some theoretical description of the ability being tested is needed that facilitates item development.

Second, items must be generated in a manner consistent with the content domain. How does one convert abstract statements of content into test items. This is the aspect of item writing which is seldom treated and is generally not a well known technology of testing.

**Table 3**

**A Validated Set of Multiple-Choice Item Writing Rules<sup>1</sup>**

General Item-Writing (Procedural)

1. Use either the best answer or the correct answer format.
2. Avoid complex multiple-choice (Type K) items.
3. Format the item vertically not horizontally.
4. Allow time for editing and other types of item revisions.
5. Use good grammar, punctuation, and spelling consistently.
6. Minimize examinee reading time in phrasing each item.
7. Avoid trick items, those which mislead or deceive examinees into answering incorrectly.

General Item Writing (Content Concerns)

8. Base each item on an educational or instructional objective.
9. Focus on a single problem.
10. Keep the vocabulary consistent with the examinee's level of understanding.
11. Avoid cuing one item with another; keep items independent of one another.
12. Use the author's examples as a basis for developing your items.
13. Avoid overspecific knowledge when developing the item.
14. Avoid textbook, verbatim phrasing when developing the item.
15. Avoid items based on opinions.
16. Use multiple-choice to measure higher level thinking.
17. Test for important or significant material; avoid trivial material.

Stem Construction

18. State the stem in question form.
19. When using the completion format, don't leave a blank for completion in the beginning or middle of the stem.
20. Ensure that the directions in the stem are clear, and that wording lets the examinee know exactly what is being asked.
21. Avoid window dressing (excessive verbiage) in the stem.
22. Word the stem positively; avoid negative phrasing.
23. Include the central idea and most of the phrasing in the stem.

General Option Development

24. Use as many good distractors as are feasible; more options are desirable.

25. Place options in logical or numerical order.
26. Keep options independent; options should not be overlapping.
27. Keep all options in an item homogeneous in content.
28. Keep the length of options fairly consistent.
29. Avoid, or use sparingly, the phrase "all of the above."
30. Avoid, or use sparingly, the phrase "none of the above."
31. Avoid the use of the phrase "I don't know."
32. Phrase options positively, not negatively.
33. Avoid distractors that can clue testwise examinees; for example, avoid clang associations, absurd options, formal prompts, or semantic (overly specific or overly general) clues.
34. Avoid giving clues through the use of faulty grammatical construction.
35. Avoid specific determiners, such as "never" and "always."

#### Correct Option Development

36. Position the correct option so that it appears about the same number of times in each possible position for a set of items.
37. Make sure there is one and only one correct option.

#### Distractor Development

38. Use plausible distractors; avoid illogical distractors.
39. Incorporate common errors of students in distractors.
40. Avoid technically phrased distractors.
41. Use familiar yet incorrect phrases as distractors.
42. Use true statements that do not correctly answer the item.
43. Avoid the use of humor when developing options.



Third, content experts must read each item and determine if the item measures the content it is supposed to measure. This activity is generally one which forces the content expert to classify the item according to content (fact, concept, principle, or procedure) and a level of cognitive behavior, such as represented in Bloom's taxonomy. Not surprisingly, content experts will disagree about the classification of an item, perhaps because of the ambiguity of the method used to generate the item or perhaps due to the possibility that an item may be multidimensional in content (see Reckase, 1985). Nonetheless, the content classification of the item is essential, because later when tests are built, the test construction must follow content specifications strictly.

Fourth, the content experts must verify that the keyed answer is indeed correct. In certification and licensing testing, this is virtually a requirement, because items must have an authoritative reference to verify its correctness. In litigation against a certification board, faulty items may provide a valid legal basis for challenging examination results. Interestingly, in virtually any area of testing, content experts will invariably disagree on a correct answer on certain items. Such items should best be revised until the experts agree or retired from use.

The content review is an extensive and important aspect of item evaluation. As personnel testing and certification and licensure testing continue to provide information for important decisions, it is critical that test users ensure that the Standards for Educational and Psychological Testing (1985) are followed. The potential adverse impact of weak examinations due to content problems, both in terms of the public and the examinees, is great.

### Editorial Review

Face validity is an important consideration in any testing situation. Test-takers should respect a test and the test makers. One characteristic of any test is its production. Tests should look like tests. Test items should be clearly presented, well written, and free of annoying grammatical, spelling, capitalization and punctuation errors. Therefore, every test and test item should be subjected to an editorial review.

In "serious" testing programs, which often involve many examinees, it is typical to employ editors who are trained and experienced in technical writing, with emphasis on testing. One aid in editorial review is an editorial guide, which contains the rules and regulations governing how items are written. The editorial guide provides a syllabus of conventions, usable acronyms, writing style guidelines, acceptable formats, and other directions, which ensure editorial consistency within the test.

With less formal testing programs, the function of an editorial review can be done by a person who is skillful in the use of the written English language.

There are a plethora of minutiae in the editorial review, with which the casual test user generally is unconcerned or unaware. Nonetheless these details of test production, which fall into this category of editorial review, are important in conveying to the test user and test consumer that the test is a legitimate one. Some of these tedious details are (1) correct alignment of items, (2) the correct use of option designations (ABCDE), (3) correct numbering of items and page numbers, (4) clear instructions, (5) readable print font and type size, and (6) proper reference to figures, charts, and photographs in the text of the test.

There are a number of commercial products, text analyzers, that will provide a measure of the reading level of the test items, whether or not an active or passive voice is used (active is preferred), and other grammatical principles and practices. These are valuable aids to the technical editor in this editorial review.

### Bias Review

The civil rights movement coupled with Uniform Guidelines, the Standards for Educational and Psychological Testing, Code of Fair Testing Practices, and increasing litigation over the appropriate use of testing has made test makers more mindful of the threat of bias in testing. The fact that important decisions are made about individuals based on test results only increases legal monitoring of testing practices and abuses of test results. Another factor that persists in all of ability and achievement testing is that test scores of minorities are typically lower than the majority culture, and that males consistently outscore females. The issue here is an unsolvable dilemma: Are these differences real? or the result of a biased testing system? While not addressing this profound problem directly, we test makers must ensure that all tests are free of any bias that would discriminate unfairly against women and minorities.

Generally, there are two interpretations of bias. One is simply that tests produce artificial differences in subgroups; differences that are relevant to the achievement domain or ability being measured but something indigenous to these group differences. A second interpretation of bias is the issue of a fair use of a test result, for instance, to use a vocabulary test to hire an English teacher.

The study of bias of test items is complex (see Cole and Moss, 1988). There are a number of prevailing statistical methods for studying item bias in objectively scorable tests. With large testing programs, it is critical to perform such studies, as part of validation research. In informal or small testing programs, such studies are difficult or virtually impossible to conduct.

Another kind of bias review is to have minority representatives examine items for culturally loaded bias. The reference to snowshoes in a standardized achievement test may be odd to persons who grew up and lived in the Sonoran desert of Arizona. This kind of test item represents a regional bias. A question dealing with how to compute a baseball batting average may be biased against



women. Item biases that affect various ethnic groups in America may be more subtle to identify but nonetheless bear close scrutiny.

### Evaluation of Distractors

Only recently has attention been drawn to the evaluation of multiple-choice distractors (Thissen, Steinberg, and Fitzpatrick, 1989; Wainer, 1989). Curiously, interest in using responses to distractors to score tests was initiated in the 1930s (see a review by Haladyna and Simpson, 1988), but research on the scoring of wrong answers has not been convincing. This may be partly due to the fact that items are not particularly well written for this kind of scoring. Another limitation is that studies done several decades ago were poorly conceived or lacked a validity context. Still another limitation is that methods for scoring wrong answers were crude and insensitive.

With the advent of item response theory, such theorists as Bock (1971), Levine and Drasgow (1987), Samejima (1979), Simpson (1983; 1986), Thissen and Steinberg (1984), and Wainer and Kiely (1987) has suggested ways to study distractor performance and to use test results in interesting and important ways, to score tests, to detect aberrant response patterns, and to evaluate distractors. This section will limit itself to the latter concern.

There is a traditional method for evaluating distractors (Millman and Greene, 1988). Generally, examinee responses to distractors must be negatively correlated to total test score and be selected by examinees at a minimum 5% of the items. Of course, those selecting a distractor should be exclusively low-scoring examinees, if the distractor is working as intended. Unfortunately, there is little research where this traditional manner of evaluating distractors is used. Haladyna and Downing (1988) examined several standardized examinations using this traditional evaluation criteria and discovered that the typical multiple-choice item had only one or two working distractors. They concluded on the basis of their study and other studies on the desirable number of options, that the three-option item is probably sufficient for most testing purposes. In other words, the typical item writer only produces one or two really good distractors when writing items. The use of useless third and fourth distractors in a four- or five-option item seems pointless. Moreover it makes takes more time to write such items and administer tests consisting of these items. If these results are confirmed by other studies, then test makers should concentrate more on distractors and write fewer, better distractors per item. The standard evaluation technique, outlined by Millman and Greene (1988) provides a useful, familiar tool for evaluating distractors, one that is compatible with your standard item analysis program.

A more recent approach to evaluating distractors derives from item response theory, but in actuality is not based on any theory. Thissen, Steinberg, and Fitzpatrick (1989) suggest the use of option graphs called "trace lines" to provide an picture of item performance. From item response theory, we have learned that a trace line should increase as a function of the underlying trait (achievement or ability) being measured. Correspondingly, a trace line for a distractor should decrease. Figure 1 shows a variety of trace lines for a single item. Option A is a correct answer and displays a classical monotonically increasing trace line. Option B is a distractor and displays a classical monotonically decreasing trace line. Option C displays a non monotonic trace line, which suggests that the option discriminates with respect middle-scoring examinees. This information can be used in several polychotomous item response theory models (Simpson, 1983; 1986; Thissen and Steinberg, 1984). Option D is a low-response option, suggesting that it is so implausible that only an occasional random guesser would choose it. These are the kinds of distractors that probably should be eliminated from tests, resulting in shorter tests with shorter administration times. The removal of useless distractors also permits the use of more test items within the constraints of a time limit, hence better sampling of the content domain.

### Summary

This paper has sought to clarify and review elements of any testing program related to item development and evaluation. As statistical test theory continues to make incredible advances, the science of item writing is somewhat emerging but hardly keeping pace. Test developers should continue to apply emerging technologies in item writing to improve tests. New formats and methods offer ways to build better items, hence better tests. Research is also needed on these emerging methods. Traditional item analysis is being replaced by item evaluation which has many important facets. As we realize that the item is the basic building block of a test, the more work we put into each item in terms of its development and evaluation, the better our tests will be.

### REFERENCES

- Albanese, M. (1990). The type K item. Paper presented at the symposium "Expanding the symposium for test item writing" at the annual meeting of the National Council on Measurement in Education. Boston, MA.
- American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37, 29-51.
- Bormuth, J. R. (1970). On the theory of achievement test items. Chicago, IL: University of Chicago Press.

- Cohen, S. A. (1987). Instructional alignment: Searching for the magic bullet. Educational Researcher, 16, 16-20.
- Cole, N. S. & Moss, P. A. (1988). Bias in test use. In R. L. Linn (Ed.) Educational Measurement (3rd ed.). Washington, DC: American Council on Education, MacMillan Publishing, pp. 221-262.
- Cronbach, L. J. (1970). Review of 'On the theory of achievement test items' by J. R. Bormuth. Psychometrika, 35, 509-511.
- Downing, S. M. (1990). True-false and alternate-choice formats: A review of research. Paper presented at the symposium "Expanding the symposium for test item writing" at the annual meeting of the National Council on Measurement in Education. Boston, MA.
- Ebel, R. L. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Fitzpatrick, A. R. (1983). The meaning of content validity. Applied Psychological Measurement, 7, 3-13.
- Frisbie, D. A. (1990). The evolution of the multiple true-false format. Paper presented at the symposium "Expanding the symposium for test item writing" at the annual meeting of the National Council on Measurement in Education. Boston, MA.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519-521.
- Guttman, L. (1969). Integration of test design and analysis. Proceedings of the 1969 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.
- Haladyna, T. M. (1990). Context-dependent item sets. Paper presented at the symposium "Expanding the symposium for test item writing" at the annual meeting of the National Council on Measurement in Education. Boston, MA.
- Haladyna, T. M. (1989). Generic questioning strategies for the teaching of statistics. Paper presented at the annual meeting of the Arizona Educational Research Organization. Mesa, AZ.
- Haladyna, T. M. (1987). Three components in the establishment of a certification testing program. Evaluation in the Health Professions, 10, 139-172.
- Haladyna, T. M. (in press). Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. Applied Measurement in Education.
- Haladyna, T. M. & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 1, 37-50.
- Haladyna, T. M. & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 1, 51-78.
- Haladyna, T. M. & Downing, S. M. (1988). Functional distractors: Implications for test-item writing and test design. A paper presented at the annual meeting of the National Council on Measurement in Education.
- Haladyna, T. M. & Shindoll, R. R. (1989). Item shells: A new method for writing multiple-choice test items. Evaluation for the Health Professions, 12(1), 97-106.
- Haladyna, T. M. & Simpson, J. B. (1988). Empirically based polychotomous scoring of multiple-choice test items: A review. Paper presented in the symposium "New Development in Polychotomous Scoring" at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Hively, W. (1974). Introduction to domain-referenced testing. Educational Technology, 14, 5-10.
- Levine, M. V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validation studies. Educational and Psychological Measurement, 35, 42-56.
- Lord, F. M. (1977). Optimal number of choices per item--a comparison of four approaches. Journal of Educational Measurement, 14, 33-38.
- Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. New York, NY: McGraw-Hill.
- Miller, H. G., Williams, R. G., & Haladyna, T. M. (1978). Beyond facts: Objective ways to measure thinking. Englewood Cliffs, NJ: Educational Technology Publications.

- Millman, J. & Greene, J. (1988). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.) Educational Measurement (3rd ed.). Washington, DC: American Council on Education, p. 335-366.
- Nickerson, J. S. (1989). New directions in educational assessment, Educational Researcher, 18, 3-7.
- Nitko, A. J. (1984). Book review of Roid and Haladyna's *A technology for test-item writing*. Journal of Educational Measurement, 21, 210-204.
- Nitko, A. J. (1988). Integrating teaching and testing. In R. L. Linn (Ed.) Educational Measurement (3rd ed.). Washington, DC: American Council on Education, p.447-474.
- Reckase, M. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.
- Roid, G.H. & Haladyna, T.M. (1982). A technology for test-item writing. New York, NY: Academic Press.
- Roid, G.H. & Haladyna, T.M. (1978). The use of domains and item forms in the formative evaluation of instructional materials. Educational and Psychological Measurement, 38, 19-28.
- Roid, G.H. & Haladyna, T.M. (1980). Toward a technology of test item writing. Review of Educational Research, 50, 293-314.
- Samejima, F. (1979). A new family of models for the multiple-choice item. Office of Naval Research Report 79-4. Knoxville, TN: The University of Tennessee.
- Seddon, G. M. (1976). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. Review of Educational Research, 48, 303-323.
- Snow, R. E. & Lohman, D. F. (1988). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.) Educational Measurement (3rd ed.). Washington, DC: American Council on Education, pp. 263-332.
- Simpson, J. B. (1983). A new item response theory model for calibrating multiple-choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.
- Simpson, J. B. (1986). Extracting information from wrong answers in computerized adaptive testing. In B. F. Green (Chair), New Developments in computerized adaptive testing. Symposium conducted at the annual meeting of the American Psychological Association, Washington, DC.
- Thissen, D. & Steinberg (1984). A response model for multiple-choice items. Psychometrika, 49, 501-519.
- Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement, 26(2), 161-175.
- Tiemann, P. W. & Markle, S. M. (1978). Analyzing instructional content: A guide to instruction and evaluation. Champaign, IL: Stipes Publishing Company.
- Wainer, H. (1989). The future of item analysis. Journal of Educational Measurement, 26(2), 191-208.
- Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 95-202.
- Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.) Educational Measurement (2nd ed.). Washington, DC: American Council on Education, pp. 81-129.

# Trace Lines

## Four-Option Multiple Choice

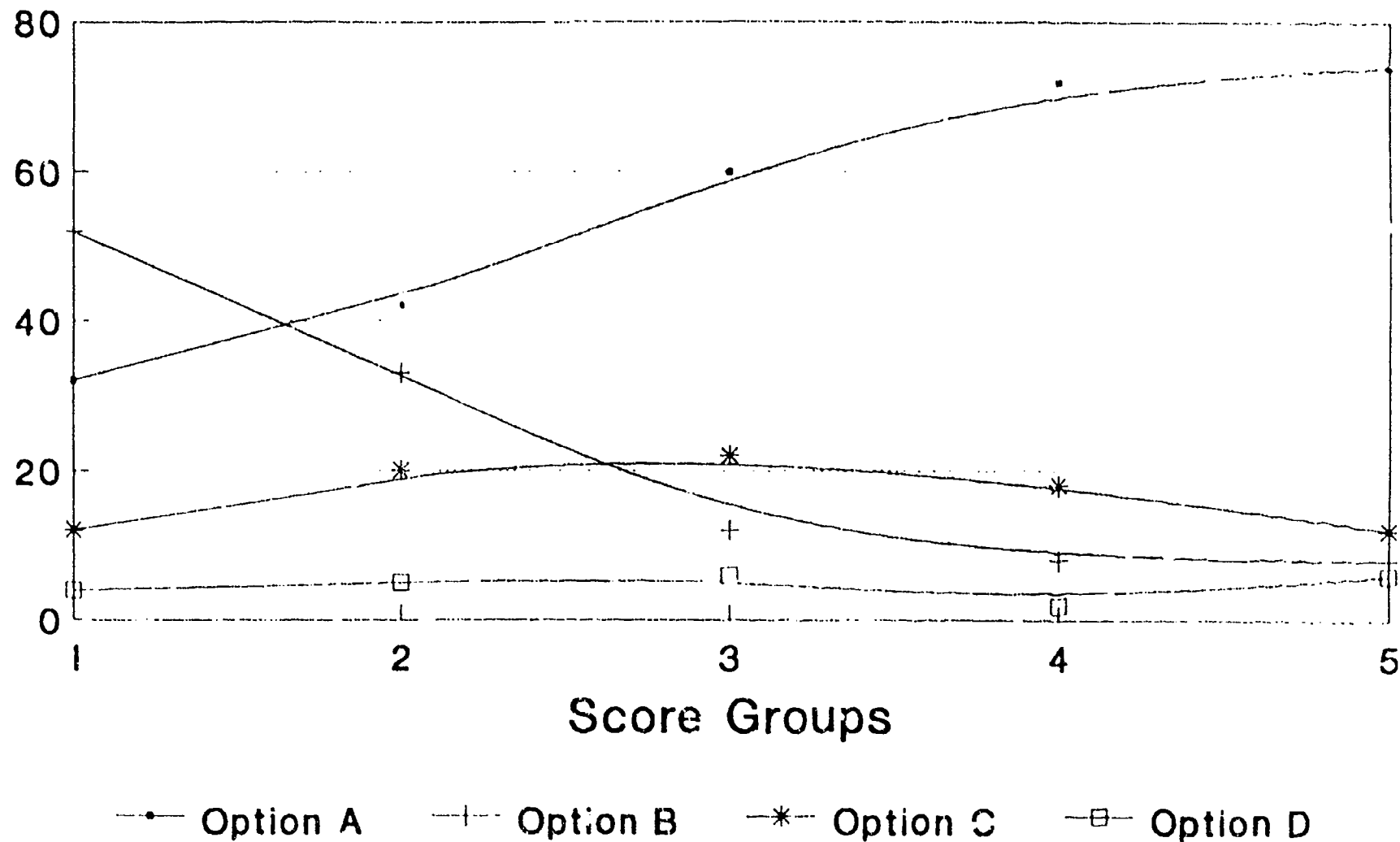


Figure 1. Trace lines for a Multiple-Choice Test Item

**LEADERS FOR WORKFORCE 2000: INNOVATIVE STRATEGIES FOR  
MEETING SELECTION NEEDS**

**Paper presented at the IPMA Assessment Council  
1990 Annual Conference on Personnel Assessment**

**Jamie J. Carlyle  
Carol A. Hayashida**

## INTRODUCTION

It has been predicted, by such groups as the Hudson Institute in their 1987 report *Workforce 2000*, that with shrinking labor pools in the year 2000, organizations will face fierce competition for skilled, qualified workers. Recent legislative proposals on Federal pay reform have highlighted the problems that the Federal sector now faces in trying to recruit and retain qualified employees--and, regardless of the outcome of these legislative proposals, things will probably only get tougher. Increasingly, budget constraints and reductions often equate to hiring or promotion freezes, which ultimately result in fewer people doing the same amount of (or more) work. This, coupled with predictions that citizens will be wanting more, and not less, service from the Government in the future will mean increased workloads for Government agencies. Such challenges indicate that Federal agencies (and non-Federal organizations as well) will be in a mode of "do more with less" as they enter the year 2000. Will we be prepared to manage in such a world?

Organizations which will be able to capitalize on these predictions of significant changes are those that become pro-active in planning for those changes. A key factor in this planning involves determining whether the organization will have the kind of leadership needed to rise to the challenge of a changing work environment. While the quality of top level management will continue to be important, the role of lower level managers, specifically first-line supervisors, will become even more critical to organizations in the future. The first-line supervisor is the vital link between employees and upper management. Not only is the supervisor responsible for assigning, directing, and evaluating subordinates' work, but as the link between employees and upper management, the supervisor must communicate the organization's mission and goals to employees in a meaningful way that will motivate them toward effective performance. The most challenging aspect of this position may be that, even though the supervisor is familiar with the technical work of the unit, the supervisory duties which are so critical to the organization's performance are often completely foreign to new first-line supervisors. Thus, the part of the job which is most critical from an organizational standpoint is also the most difficult for a new supervisor to perform.

If predictions by such groups as the Hudson Institute hold true, we will begin to see jobs increasing in technical complexity while the skill levels of individuals entering those jobs are declining (compared to present levels of entry workers). Thus, first-line supervisors may increasingly find themselves in a coaching/teaching role. Given that many first-line supervisors find the



"supervisory" aspects of their jobs difficult now, what will this mean when the need for such skills is greater and even more essential to organizational success? Are we doing anything now to ensure that our leaders of tomorrow will be able to meet these challenges?

### PRACTICES IN THE FEDERAL SECTOR

Well, as far as the Federal sector is concerned, the answer would probably be a qualified "no". A recent Merit Systems Protection Board (MSPB) study looked at how Federal agencies were preparing for the future (and dealing with the present) in terms of selecting employees who would fill those critical first-line supervisory positions. What we found was that, by and large, most Federal agencies treated first-line supervisory positions much like their non-supervisory positions, in terms of assessment for selection purposes. Not surprisingly, we also found that many of these agencies found their selection techniques to be wanting in this area. However, we were very pleased to find a few bright spots in this area. A number of agencies are trying some fairly creative approaches to assessment and selection. This paper will focus on those innovative Federal approaches and how they might assist organizations in meeting special selection needs that we'll all be facing in the year 2000.

Before describing some of these more creative approaches, it is useful to note our study findings concerning the "typical" approach used to select first-line supervisors in Federal agencies today. We asked the 22 largest Federal agencies to provide us any information they had about any methods, techniques, forms or systems they currently used for selecting white collar first-line supervisors. Eighteen agencies provided information in response to our request. From this information, it was clear that most agencies use the same general approach to supervisory selection. We selected one "prototype" agency using this typical approach, as well as four other agencies using more creative approaches to supervisory selection for further study. We interviewed both supervisors selected under these systems and their superiors concerning their impressions of the systems. We also administered written questionnaires to the subordinates of these first-line supervisors to determine their perceptions of the supervisors' performance. In addition, appraisal data were gathered from personnel records concerning the performance of these first-line supervisors.

From this, we found that the "typical" approach to supervisory selection has these features:

- Some type of job analysis is performed to identify job requirements;

- An individual job announcement is posted for each job as it becomes vacant;
- Interested individuals must submit an application form for each job as it's announced;
- Determination of best qualified applicants is based on evaluation of training and experience almost exclusively through written documentation;
- When interviews are used in the process they are of the selection rather than evaluation variety (i.e., only the few best qualified are interviewed, questions are not structured enough to be scored in any systematic, reliable way), and;
- Technical, rather than "supervisory" skills and abilities are typically emphasized in the assessment process (although there is much similarity in the "supervisory" factors that are used).

The information we collected from agencies, as well as findings from other studies which have examined supervisory effectiveness led us to conclude that the selection strategy used by most Federal agencies (i.e., one which relies primarily on the evaluation of previous training and experience) may not be adequate for meeting selection needs in all situations, and most certainly will be found lacking as we begin to see some of the changes in the workforce as predicted for the year 2000. With that in mind, some of the innovative Federal strategies are described below, within the context of how these strategies might meet the challenges facing us in the year 2000.

#### CHALLENGES FACED IN THE YEAR 2000

*Challenge #1: Scarcity of technical skills needed to do increasingly complex work.*

As mentioned previously, there have been predictions by the Hudson Institute and others that there will be an increasing need for technical skills and capabilities in the fast-growth jobs of the future. Unfortunately, many of the workers who will comprise the bulk of new hires are those who have traditionally been educationally disadvantaged and in low skill occupations. Because of this "collision course" looming between projected needs and projected intake, the supervisor of the year 2000 will need to stay on top of the state-of-the-art in his or her technical field. New hires cannot be expected to be well-versed in advances in a technical field in which they may be only marginally

competent; thus, supervisors may need to be able to perform on-the-job teaching or coaching in the technical area.

While the "typical" selection method may be sufficient for assessing technical knowledge, what is needed is a system which can more efficiently assess both supervisory and technical qualifications at the same time. One agency's program that we studied, the U. S. Marshals Service (USMS), employs as one component of its evaluation strategy, a written knowledge test which measures candidates' knowledge concerning both technical and supervisory aspects of the job. This test, which is part of the process to select candidates for Supervisory Deputy Marshal jobs, is designed to assess a candidate's knowledge of laws, regulations, processes, and operating procedures necessary to supervise the work performed by Deputy Marshals. The advantages of such tests for agencies are numerous: they are relatively easy to administer and score (especially as compared to procedures used to evaluate training and experience documentation); applicants generally perceive such measures as fairer than more subjective assessments, and; agencies can use data gathered from administrations of these tests over time to determine future training and development needs.

One of the major drawbacks to the method, of course, is that it rests on the assumption that an applicant would actually apply the knowledge that he or she possesses on the job; however, by the same token, traditional methods relying on evaluation of training and experience also assume that exposure to certain training and experience will result in the application of what was learned. Therefore, in that sense, there is no difference.

Another system which we studied which provides an interesting solution to the need to ensure a certain level of technical skill in newly selected supervisors is that found in the Department of Labor's Mine Safety and Health Administration (MSHA). The strategy developed by MSHA requires all those selected as candidates for Supervisory Coal Mine Inspector positions to undergo an intensive, one-year training program prior to placement in the supervisory position. The approach MSHA uses includes training in specific technical topics as well as training in supervisory skills and subjects. The program uses centrally administered, formal classroom sessions, as well as short-term job assignments in which participants work with an incumbent first-line supervisor. Throughout the course of the program candidates are formally tested to ensure they are acquiring the desired level of knowledges and skills.

Although such a program requires a significant commitment of resources to develop and administer, it could prove economically worthwhile for an organization in the long run

if the potential cost of "mistakes" made by supervisors lacking in technical knowledges and abilities were substantial. An approach like MSHA's program also offers the advantage of "standardized" training, which may be important for some organizations.

*Challenge #2: Need to increase productivity in order to efficiently use increasingly diminishing resources.*

Over the last several years there has been a growing concern over the nation's "productivity problem." And while no facet of American industry has escaped the problem, it becomes especially significant for the service industries because so much of our nation's future lies in the these industries. Productivity gains in the service industries (including Federal agencies) will be critical to economic growth in this country. We won't be able to bolster productivity by employing more workers because fewer people are expected to be entering the workforce. Therefore, other approaches will be needed to ensure that competitive levels of productivity are achieved.

In the last several decades we've witnessed a proliferation of "fixes" to this productivity problem. One that has remained popular, however, is the notion of participative management. Many organizations are trying to foster climates of participative management by involving employees in decision making concerning various aspects of their work lives. In one organization that we studied, decisions concerning supervisory selections have become a primary focus for employee participation.

The typical procedure for assessing candidates for supervisory positions offers little or no opportunity for subordinates to have a voice in the decision made. However, the Federal Aviation Administration (FAA), as part of a larger supervisory identification and development program, has implemented a peer rating strategy which enables subordinates to have input into the assessment process. In this particular system, applicants for FAA's Supervisory Air Traffic Controller jobs are given peer ratings (i.e., ratings made by other Air Traffic Controllers with whom they work) on supervisory skills and abilities. The elements rated include (but aren't limited to) interpersonal and communication skills and abilities. Users whom we surveyed in our study agreed that this strategy can provide unique insights concerning applicants' potential to perform the supervisory job. Because the interpersonal skills involve interaction with others, a strategy which is based on the perceptions of those who have the greatest opportunity to interact with the applicant on a daily basis provides a very direct measure of those skills and abilities.



When employees are able to provide input concerning applicants' qualifications for the job, they may be more likely to accept selections subsequently made. Fostering employee participation through peer ratings also helps to communicate to employees that their input is important to management, which can enhance the work environment, and, hopefully, productivity. Such a strategy might have the opposite effect, however, if employees are given the opportunity to provide input, but the input is not used in making the selection decision (e.g., a candidate who is consistently rated poorly by peers is nonetheless selected for the job).

### *Challenge #3: Dramatic changes in workforce demographics.*

It has been predicted that by the year 2000 approximately 47% of the workforce will be women, and 15% of the workforce will be minorities. What will this changing workforce composition mean for supervisory selection? Well, for one thing, methods (such as that typically found in the Federal sector) which rely on evaluation of previous training and experience may not be compatible with the realities of a changing workforce. That is, when many occupational fields have historically been filled by white males, women and minorities desiring to move up into supervisory positions may not have had the opportunities to acquire the particular training and experience used to infer possession of the necessary knowledges, skills, and abilities to perform the supervisory job. Thus, if training and experience measures were continued to be used, there would be even greater disparities (in terms of sex and racial/ethnic representation) between employee and management segments of the workforce. Therefore, a strategy is needed in which the evaluation of applicants' qualifications is based on actual demonstrations of performance, rather than evidence of previous training and experience. Several of the programs we studied have employed such strategies.

Both the FAA and the USMS use simulation exercises as part of their processes to evaluate applicants for supervisory positions. With these simulation exercises, applicants are presented with scenarios depicting situations typically encountered in the supervisory job. Applicants must "size up" the situations presented, articulate the issues or problems involved, and take whatever actions they feel the situations call for.

The simulation exercises are designed to enable the applicants to demonstrate performance on supervisory abilities such as oral communication, decision-making, and leadership. They rely on candidates' "on-the-spot" performance relevant to these abilities. Therefore, candidates aren't penalized in the evaluation process by a

lack of prior opportunities to demonstrate their qualifications through work experience or training. This enables female and minority candidates to be evaluated along with male, nonminority candidates fairly and effectively, based upon present performance.

Simulation exercises have been used in different ways to enhance the evaluation process. For example, the FAA (in their supervisory identification and development program) uses simulation exercises as part of a skill-based interview, while the USMS uses them in an assessment center to select first-line supervisors. One primary advantage to the use of simulation exercises (in addition to alleviating biases associated with lack of opportunity to gain certain training and work experience) concerns the feedback that is typically provided by this method. Candidates participating in these exercises usually receive very detailed feedback concerning the strengths and weaknesses in their performance, which can prove very useful for developmental purposes. This step also gives a boost to the non-selectees and sets the stage for personal and professional growth that can enable them to be more competitive candidates the next time they attempt to move into the supervisory ranks.

One of the major disadvantages to the use of simulation exercises, however, is the cost involved in their development and administration. In order to provide reliable and valid predictions concerning job performance, the exercises must be realistic to candidates and assessors. The development of such exercises requires extensive involvement of subject matter experts, and evaluators used in the process must be thoroughly trained in specific procedures used to assess the candidates. However, despite the commitment of resources required, most of the candidates, selecting officials, and administrators we spoke with in our study view these exercises as a much more valuable tool for enabling candidates to demonstrate their potential for supervisory positions than traditional training and experience ratings.

#### *Challenge #4: Desire to reduce turnover in the workforce.*

Because of the projected labor shortages in the workforce by the year 2000, there will be fierce competition for qualified employees. Unfortunately, this competition may translate to high turnover among employees for many organizations. Also, because of predicted changes in the composition of the workforce (e.g., more women, an aging workforce population), there will also likely be increased demands for creative work programs to meet the needs of these workers (e.g., flexiplace, job-sharing). Organizations not prepared to meet these demands face stiff competition from (and, thus, higher turnover than) those



organizations that are prepared to provide effective responses to their employees' personal needs.

Therefore, high employee turnover may become a threat which hangs heavily over many organizations by the year 2000. An organization which takes steps necessary to prevent unnecessarily high turnover will surely be more competitive than those not taking such steps. One can see the critical role first-line supervisors can play in this endeavor by imagining how bad supervision might contribute to turnover. (In fact, a recent MSPB study found quality of supervision to be a factor in many employees' decisions to leave their Federal jobs.) An ineffective or poor supervisor can wreak havoc among employees, leading to frustration, dissatisfaction, and eventually turnover. First-line supervisors who are able to "hit the ground running" in their jobs and are equipped to handle their personnel management roles (including knowing enough about the personnel management system to effectively respond creatively to subordinates' needs) have an advantage in terms of preventing turnover over those not so prepared.

Thus, an assessment approach is needed which can ensure that supervisors will be able to "hit the ground running" (i.e., any supervisor selected will have at least a minimum level of competency in or knowledge of particular supervisory functions). Two methods previously discussed can help agencies meet this need. These are the written tests (objectively scored measures of possession of certain technical and supervisory knowledges), and pre-placement training and evaluation (e.g., MSHA's supervisory pool program). Depending on an organization's particular mission and employee needs, tests or pre-placement training and evaluation programs could be developed which focus on those knowledges, skills, and abilities seen as most critical for ensuring that new first-line supervisors can be effective from the very beginning, with very little "downtime" required for learning the job. As noted previously, there are advantages and disadvantages to both of these strategies; nonetheless, in terms of meeting this particular challenge, such strategies appear to offer some practical solutions.

#### *Challenge #5: Increased reliance on automation.*

It almost goes without saying that we're becoming more and more reliant on automation in all aspects of our lives, but especially in the work place. There is every reason to believe that this trend will continue, and in the area of evaluation and selection, it will become more and more critical as timeliness and efficiency become important factors in competing for scarce resources. Automation of assessment and selection can have an impact not only on the

speed and efficiency of the process, but also on applicants' perceptions of the process. That is, by streamlining the application process, applicants may be more likely to apply for positions, thus increasing the organization's competitiveness for human resources.

The typical selection procedure employed in the Federal Government today uses job analysis to identify relevant skills or knowledges for individual jobs and announces vacancies for supervisory positions as they occur. Candidates interested in these positions must submit a separate application package (i.e., documentation concerning job qualifications) for each position. Depending on the number of applicants and the factors being evaluated this strategy may be done in a timely and efficient manner by some organizations today, but, by and large it's seen as cumbersome, slow, paper intensive, and one which discourages many potentially qualified candidates from applying for positions in the first place. As organizations are forced to become more competitive (and with shrinking budgets, also looking for ways to make their selection processes more efficient), methods which enable the automation of the process become more and more attractive.

One of the systems we studied which has effectively automated much of this process is the Department of the Army's Army Civilian Career Evaluation System (ACCES). This system uses standardized evaluation criteria for related jobs to fill both supervisory and nonsupervisory positions in many occupations at mid-level grades and above throughout the agency. Through extensive job analyses, they identified inventories of evaluation criteria (knowledges, skills, abilities--KSA's) relevant to groups of related jobs (referred to as "career programs"). An employee interested in being considered for job referrals under ACCES only needs to submit one application package containing descriptions of accomplishments and self and supervisory ratings on evaluation criteria relevant to the career program. This information is stored in a central computer file, and whenever a vacancy occurs of the type and location in which the applicant is interested, the applicant is automatically considered for the position.

An applicant doesn't receive an overall referral score or ranking after submitting an application package to ACCES, because a candidate's referral score and ranking may change with every vacancy filled. This is due to two aspects of the system: 1) the KSA's are weighted according to the particular requirements of each position applied for (the applicants' referral score is based on both the weights of the rating elements and the applicants' ratings on the elements), and; 2) the mix of applicants interested in positions will vary by location of the vacancies.

Of course, in addition to ACCES, other selection systems we studied have components that can be (and in some cases have been) automated. For example, scores generated in the USMS's written tests and assessment centers are stored in automated files so that candidates can be considered automatically when positions become vacant. The next logical step (which we will likely see more of in the future) is to eliminate much of the paper and pencil aspects of evaluation and go directly to automated data entry (e.g., written tests administered at computer terminals; interactive simulation exercises in which the candidates must choose the appropriate course of action at various points in the simulation, etc.). The possibilities are endless. The primary advantages of such strategies of course are in their efficiency and timeliness. Another primary benefit, at least in the public sector, is that automated referral strategies enable candidates to be considered automatically for positions they otherwise may not even hear about. With many Federal agencies having large, decentralized, geographically dispersed workforces, (and with a push for more decentralization in the future), this kind of strategy should become more and popular as time goes on.

### CONCLUSIONS

The challenges listed above are but a few of the many that will be facing measurement specialists as we approach the year 2000. The central theme of all the projections, however, seems to be that there will be intensified competition for well-qualified workers. This makes it all the more important that organizations be able to select their leaders effectively and efficiently from among a shrinking pool of candidates. Being able to develop and administer strategies which can do that will be a top priority for measurement specialists. Fortunately, some organizations (such as those agencies mentioned in our study) are already looking towards the future, and there is much to be learned from their experience.

Hopefully, as organizations begin to plan for the future, they will take a closer look at the adequacy of their supervisory selection systems. In doing so, it might be helpful to use as a framework for examination some of the common features that seemed to characterize the successful programs that we've discussed here.

- *Top management is visibly supportive of the system.* In several programs, the highest levels of management clearly play significant roles in the design and implementation of innovative supervisory selection systems. They are involved in approving the conceptual approaches and most importantly, they view the systems

as an integral part of their human resource management plan.

- *The system meets organizational needs.* Contrary to some approaches which may be driven by administrative or procedural requirements of the personnel system, these systems leave the distinct impression that, while they exist in the context of the personnel system, the programs are created by managers to meet managers' needs. This doesn't mean that measurement specialists or personnel specialists don't play a significant role; only that they are involved in a much more active partnership than is normally the case with selection programs.
- *The system is dynamic.* Although the systems studied were in different stages of implementation, developers and administrators of each keep a close watch on the changing needs of those served by the systems, and modify the systems accordingly. Several have implemented formal procedures for obtaining feedback concerning operation of the systems. This feedback is used as an integral part of ongoing attempts to improve the systems.
- *The system uses a sound measurement approach.* The systems we studied acknowledge the criticality of the personnel management part of a supervisor's job and have developed ways to assess the requisite skills or potential in candidates. The systems emphasize comprehensive job analysis to identify required KSA's and substantial involvement of subject matter experts in the selection process.

In conclusion, based on our experiences in this study, we believe there is a need for much greater sharing among organizations concerning practices in assessment and selection. It was amazing that few agencies we spoke with seemed to know what other agencies were doing in this area. And while it's true that, as time goes on, we may find ourselves in greater competition with one another for scarce human resources, there's still much benefit for all who share their experiences with others in this area.

**MUNICIPALITY OF METROPOLITAN SEATTLE  
(SEATTLE METRO)**

**MULTIPLE HURDLE SELECTION OF  
TRANSIT SUPERVISORS:**

**A CASE STUDY IN PROMOTIONAL  
COMPETITION & SELECTION FROM A  
DIVERSE WORK FORCE**

Presented at IPMAAC Conference - San Diego, California

June, 1990

Prepared by

John Barber

Test Development Analyst



## **ABSTRACT**

The general history and structure of Seattle Metro are discussed as background to this selection process. Previous selection procedures and special issues such as affirmative action concerns and union contract provisions affecting this selection process are described. The five job specializations included in this classification and the results of the initial job analysis are then discussed. The history of this selection process over a period of 18 months, including a concurrent validity study of two commercial tests, is then described and the structure of the four phases of the process (work record, written test, role plays, and panel interview) are discussed in detail. The results of the process are then described, indicating significant progress in affirmative action hiring without the use of race or sex conscious methods. Test results from each of the three testing formats (Written, Role Play and Oral Interview) are reviewed and their impact on groups by sex and race are described. Overall, the results suggest that it is possible to make significant progress toward achieving affirmative action goals without race and sex conscious actions. An agency must be willing, however, to use recruiting and selection methods which maximize candidate opportunities to compete in a variety of skill areas rather than screening out large numbers of candidates early in the process.

## INTRODUCTION

### Structure and History of Seattle Metro:

The Municipality of Metropolitan Seattle (known locally as METRO and nationally as Seattle Metro) is a regional local government entity which provides all public transit and sewage treatment services within the geographical area covered by the City of Seattle and the surrounding metropolitan area within King County. It is governed by a large council composed of the Mayor and City Council of the City of Seattle, the King County council and executive, and representatives from various incorporated cities and unincorporated areas in King County.

Seattle Metro is a relatively new governmental entity. It was created in 1958 by the legislature of the State of Washington and the voters of King County in order to clean up and save Lake Washington (a major fresh water lake dividing the City of Seattle from the rest of King County) by developing a sewage diversion and treatment system. The result was a nationally recognized success in saving a dying lake.

In the early 1970s, Seattle Metro was again called upon by the voters of King County to develop and operate a regional public transit system from a fragmented system of public and private transit properties. The result has been a regional transit system which has received a number of awards from national transit organizations.

In attempting to meet the needs of the citizens of King County, Seattle Metro has grown from an organization of 500 employees in 1958 to nearly 4500 in 1990. Future increases in size are projected in the next ten years as sewage treatment plants are expanded dramatically and the region considers development of rail transit systems.

Legally, Seattle Metro was created as a "merit" employer but is not governed by a full civil service system. Competitive selection processes are required in most situations but there is no requirement to use formal registers or certification procedures. Many of the rules and regulations governing personnel actions and procedures within the Transit Department are governed by the provisions of the current contract between Metro and the local transit union (local 587).

### Transit Operators and Supervisors:

Approximately 2000 of Seattle Metro's 4500 employees are part or full-time Transit Operators. About 1000 of these operators are full timers and most of those 1000 are eligible to compete periodically for promotion to a position as a first line transit supervisor.

Transit supervisors are members of the local transit union. They currently work in one of the following specialty areas:

Service Supervisor: These positions work in the field in an assigned district of the city or county where they monitor buses for adherence to schedules, safety regulations, etc. and solve problems that develop as a result of traffic problems, accidents, passenger relations problems, et They also help out with planning and operations associated with special events such as fairs, parades, etc.

Base Dispatcher/Planner: Works in a bus base assigning operators to runs, developing schedules, taking phone messages, distributing run cards and schedules, arranging coverage for last minute absences or late reports using report operators, etc.

Communications Coordinators: Works in the radio communications center, responding to calls from and directing coaches to respond to traffic conditions, etc.

Instructors:  
Instructs trainees who are about to become new transit operators, provides re-training for operators who are having problems, qualifies operators on new routes, etc.

Schedule Maker: Prepares new bus schedules based upon routes and resources prepared by the planning section.

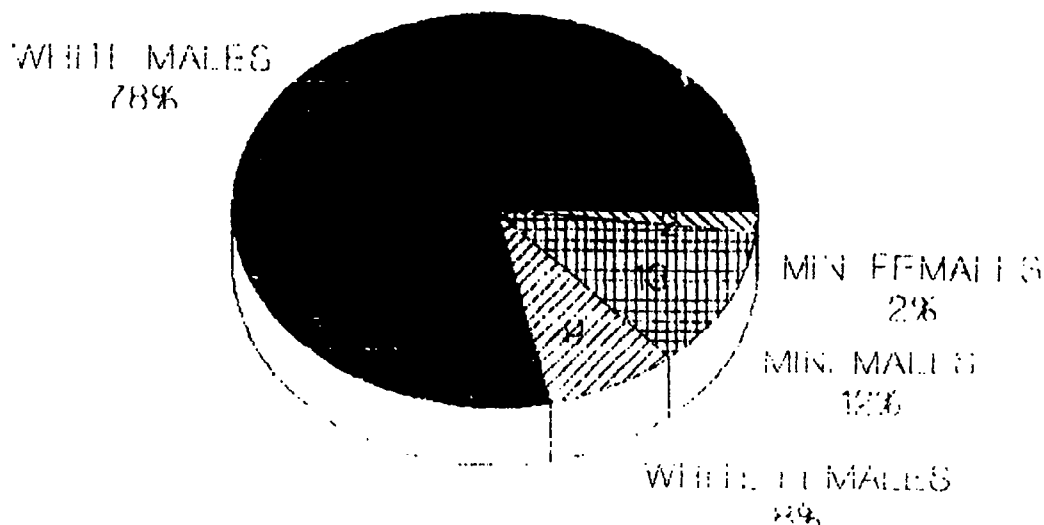
Tunnel Controller: Monitors status of alarm, security, fire protection, and monitoring systems in Seattle Metro's new transit tunnel under downtown Seattle. This is a new assignment which just started in 1990.

After their initial hiring, new transit supervisors spend a year as a "Supervisor-In-Training". During this period they are trained in and receive on-the-job experience as Service Supervisors and Base Dispatcher/Planners. Supervisors who have completed their training choose their specialty assignment from the remaining slots available after those with higher seniority choose ahead of them. The result of this system is usually that only the most senior supervisors work as Schedule Makers, Communications Coordinators and Instructors. Tunnel Coordinators will probably also be high seniority jobs in the future.

### SPECIAL SELECTION ISSUES:

This class of work is counted as belonging to the EEO category "Technicians" under Metro's affirmative action plan. Figure 1 shows the race and sex breakdown of transit supervisors as of the winter of 1988/89. As can be seen in Figure 1, females were dramatically under-represented in this group of titles.

**FIGURE 1**  
**TRANSIT SUPERVISORS BY RACE & SEX**



WINTER 1988/89

Because of the large number of transit supervisor positions, transit supervisors make up by far the largest number of technician positions in the Operations Division of the Transit Department.

Transit Supervisors at Metro are represented by the same union as part and full-time Transit Operators (Local 587 of the Amalgamated Transit Union). The provisions of the current contract between Metro and local 587 affect several aspects of the selection of Transit Supervisors.

First, the contract sets the minimum qualifications for applying at one year of current full time driving experience. Second, Metro usually uses a hiring list so that a new process is not required for each new vacancy. Under the current contract, initial hiring at the end of the selection procedure is done at Metro's discretion. All future hires must come off of a list that lasts for up to two years or until exhausted. Candidates are ranked on this list in seniority order and must be hired in that order.

#### PREVIOUS SELECTION PROCEDURES:

Up until the early 1980s, selection of transit supervisors was done using informal selection procedures such as temporary appointments and informal interviews. Then Metro began to use formal, multiple hurdle testing procedures. The last few selection processes prior to the one described here consisted of four basic components as follows: a) work record review; b) written examination; c) role plays; and d) oral interview board.

## HISTORY OF THIS SELECTION PROCESS:

### Part I - Identification of Barriers to Affirmative Hiring:

In the spring of 1988, Metro began development of its most recent selection process for transit supervisors. At that time, it was known that Metro's new transit tunnel would be ready for opening in the fall of 1990. It was also known that the tunnel opening would result in at least 20 new positions between 1990 and 1992.

Due to the large number of anticipated vacancies, the decision was made to launch an all out effort to recruit and hire affirmatively during this recruitment. At the same time the agency was concerned about the use of race or sex conscious methods to achieve this goal. Therefore the decision was made to analyze past selection procedures and attempt to identify any barriers to affirmative hiring, especially for women.

A task force consisting of representatives from the Transit Department's Operations Division and the Test Development Analyst from the Human Resources Division was formed to research past selection processes and develop the new one.

The task force began by analyzing test results from previous selection procedures. The results of this investigation indicated two major factors as barriers to affirmative hiring.

First, only about 25 women had applied out of some 200 or more applicants during the previous recruitment. It was clear that recruiting for women applicants had to be improv

Second, during the previous recruitment, hardly any women had made it past the written exam stage of the process. The task force therefore further investigated the reasons why more women had not passed the written test. It found that the final score on the previous written test was a composite formed from the actual test score and the candidate's number of years of seniority minus a certain number of points for each of various types of infractions such as late reports, unexcused absences, performance reports, accidents, etc. In addition, bonus points were awarded for candidates with perfect records. Separate analysis of each component of the scores clearly indicated that the addition of seniority points and the bonus points had the affect of giving males a substantial advantage in the final composite score.

The task force therefore undertook a two part strategy for improving affirmative hiring for women.

First, a special program for increased recruiting of women from the full time operator ranks was planned well in advance of the opening of the recruitment period. This program included "mentoring" and encouragement by supervisory personnel to try to get women and minority operators to consider applying for supervisor positions. It also included opportunities for operators to accompany Service Supervisors around during a regular shift of work and talks about the advantages of work as a supervisor by women who were already in those positions.



Second, the task force recommended a revision in the calculation of the work record and written test portions of the selection process. More specifically, it was recommended that a pass/fail criteria be developed for the work record portion and that the written test score be completely independent of the work record portion.

System wide records on all full time operators were then analyzed to aid in the development of a new pass/fail standard for the work record portion of the selection process. Data from this analysis was combined with subject matter expert judgements on satisfactory performance to develop point values for various types of infractions and a final overall pass/fail point.

#### Part II - Job Analysis & Test Development:

The five existing transit supervisor specializations had been subjected to a large scale task analysis style job analysis during the mid-1980s. The data from the earlier job analysis was combined with further job analysis conducted by the Test Development Analyst. Based upon these analyses, a preliminary plan was developed for the written test. This plan called for a three part exam as follows:

- Part I: Commercial Aptitude Tests
- Part II: Written Essay Problems
- Part III: Written Multiple Choice

The aptitude tests were recommended to attempt to measure two particular skills identified during the job analysis. One was the ability to handle a number of tasks at the same time while working under stress and distraction. The other was general fluency with language, such as would be used during interactions with operators and the public.

During the fall of 1988, Metro experienced a period of lay-offs. One of the effects of these lay-offs was a delay in the beginning of the transit supervisor selection process. As a result of this delay, the opportunity arose to conduct a concurrent validity study on the use of the commercial aptitude tests. The validity study and pre-test portion of the development process is described below.

At the same time, a critical incident type job analysis was also conducted in order to develop the problems for the essay portions of the written test. The task force subject matter experts also began work on multiple choice questions based upon Metro policies, procedures, the union contract, and various routes and runs in the Metro transit system. The latter were intended for use in an "open-book" multiple choice test which would make up Part III of the test.

The critical incident analysis was also used as a source for incidents which would later be used in the role play phase of the selection process and for situational problem type questions which were used during the final oral board portion of the process.

### Part III - Validity Research:

As mentioned above, the job analysis indicated that language fluency and the ability to handle a number of tasks at the same time under stress are important components of the job. It was also suspected that certain supervision oriented attitudes might be important to overall job performance as a supervisor.

It was decided to conduct a concurrent validity study of three commercial pre-employment tests. The tests were: a) The Press Test; b) Word Fluency and c) The Management Readiness Profile. All three tests are published by London House.

The criterion measures for the validity study were performance evaluation ratings done specifically for the validity study by the immediate supervisors of current transit supervisors in four of the five specialty areas. Performance evaluation dimensions included: a) Human Relations/Oral Communications Skills; b) Knowledge of Transit System Policies & Procedures; c) Problem Solving; d) Leadership; e) Handling Administrative Details; and f) Ability to Handle Multiple Tasks under Stress. In addition an overall rating was made by the raters on each employee and a total score was calculated by summing up the ratings on the six performance dimensions listed above.

96 current transit supervisors took the three tests described above during the winter of 1989. Their scores were then correlated with their performance evaluations on the eight criterion measures described above.

Since the average seniority of employees varies between specialty areas, the relationship between the test scores and the performance measures was analyzed separately for four of the five transit supervisor specialty areas. None of the employees in the fifth specialty area (Schedule Makers) agreed to participate in the study. Both sample size and correlations coefficients varied considerably between specialty groups. Rater reliability coefficients also varied considerably between specialty groups.

Table 1 shows the inter-correlation table for the Press Test and Word Fluency Test with the six performance evaluation dimensions and with the total points measure for the Service Supervisors group. As can be seen in Table 1, both tests showed strong positive correlations with performance ratings, especially with the handling multiple tasks dimension. Service Supervisors are the single largest specialty group and are the first group to which most new supervisors are assigned. Based upon these results, it was decided to use these two tests during the upcoming selection process.

Correlation coefficients with performance measures were not significant for any of the other three specialty areas. Analysis of rater reliability and seniority factors suggested that the lack of correlations might be due to the heavy placement of high seniority employees into these units and to problems with inter-rater reliability. The Management Readiness Profile did not correlate significantly with performance measures in any of the four specialty areas and was therefore not used during the actual testing of applicants.

TABLE 1

## Concurrent Validity Study - Correlation Matrix

Test Name	Overall Rating	Human Relations Skills	Know of Pol & Proc	Problem Solving	Leadership	Admin Detail	Mult Tasks under Strs.	Tot Pts
Press Test	.205	.254	.226	.310 *	.212	.282 *	.386 **	.316 *
Word Flu.	.269 *	.270 *	.075	.348 *	.325 *	.332 *	.442 **	.348 *

\* = P &lt; .05

\*\* = P &lt; .01

Based upon these results, a multiple regression analysis was conducted and a formula for combining the scores from the Press Test with those of the Word Fluency was developed.

During this period, pre-testing was also conducted on the written problem and multiple choice portions of the written test using employees from the Chief level in the agency. Based upon this period of pre-testing, one essay problem was dropped from the test.

#### Part IV - Work Record Review:

In March of 1989, the Transit Supervisor selection process was opened for filing. 243 full time operators applied during that time. 7 applicants did not meet the minimum qualifications and 15 failed the work record criteria.

#### Part V - Written Examination:

During April of 1989, 193 full time operators took the written test. 25 of the original applicants failed to appear for the test. The test consisted of the two commercial tests (Part I), 3 written problems (Part II), and 40 open book multiple choice items (Part III).

Administration of the test took over a week in order to schedule all applicants at a time that was not in conflict with their driving schedules.

Prior to the final grading and identification of papers, passing points based upon a satisfactory level of performance were developed for each of the three parts of the test using subject matter judgement and the results of the validity study. The passing points for each of the three parts were then combined to make a final overall passing point.

During May of 1989, the written problems and commercial tests were rated and scored by subject matters experts and Human Resources Department staff.

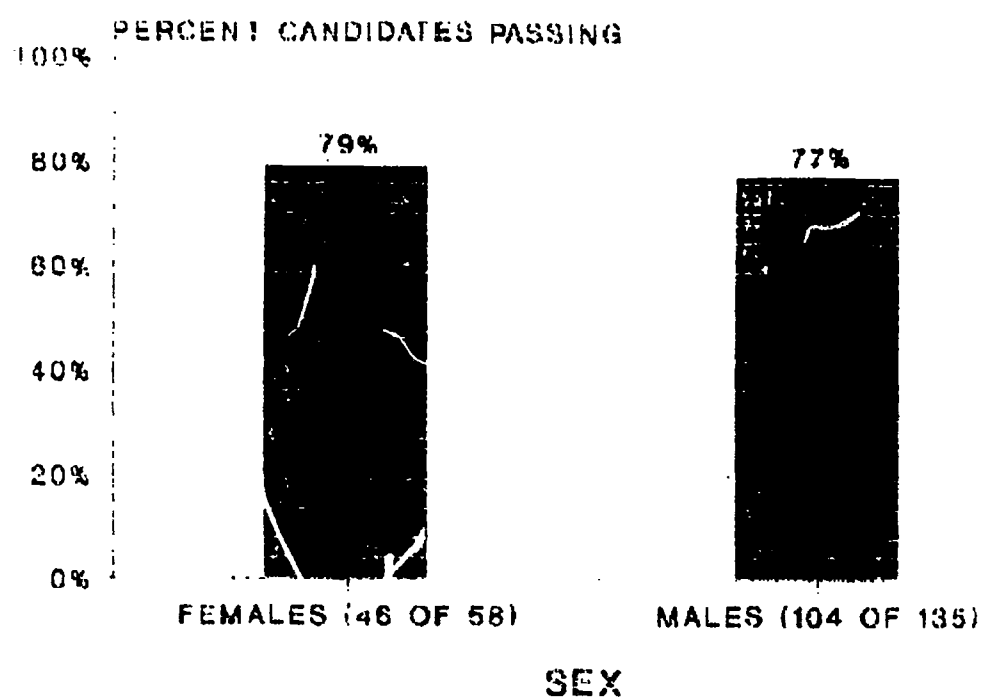
Analysis of the results of the test indicated that women had done very well but that the test had adverse impact on blacks. Two alternatives were considered for determining which candidates would continue on to the role play portion of the process. The first alternative was a top down ranking of something like the top 60 to 80 scores on the written test. The second alternative was to allow everyone who passed the satisfactory level described above to go on to the role play.

It was decided to use the satisfactory level alternative even though it meant conducting role plays for 150 applicants. Some of the factors leading to this decision were as follows:

1. Blacks were already well represented within the appropriate EEO category so race conscious actions such as within group rankings could not be defended.
2. The satisfactory level cut point could be defended as valid by virtue of being based on empirical research and subject matter expert judgement while the top down ranking method was not as defensible.
3. The satisfactory level alternative would result in many more blacks reaching the role play phase of the process.
4. Agency management decided that it was willing to invest the extra staff time and resources necessary to allow a larger group of applicants the opportunity to compete.

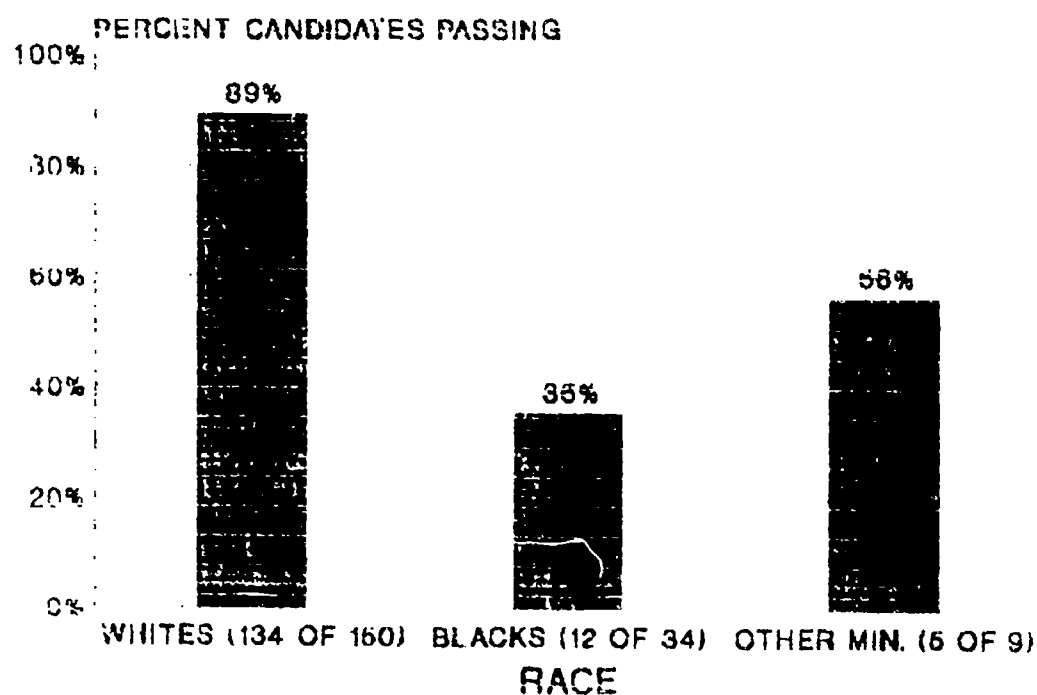
Figures 2 and 3 show the relative passing rates by sex and race respectively using the satisfactory level passing point.

**FIGURE 2**  
PERCENT CANDIDATES PASSING BY SEX



SPRING 1989

**FIGURE 3**  
PERCENT PASSING BY RACE - WRITTEN TEST



SPRING 1989



## Part VI - Role Plays:

During June of 1989, procedures for conducting role plays for 150 applicants and role plays scripts, applicant instructions and rating materials were developed. Three role play scenarios were developed. Due to the large number of applicants, it was decided to use three different rating panels. Raters were current transit supervisors and chiefs. Each rating panel was assigned a two week period during which they would observe and rate role plays for approximately 4 hours each day.

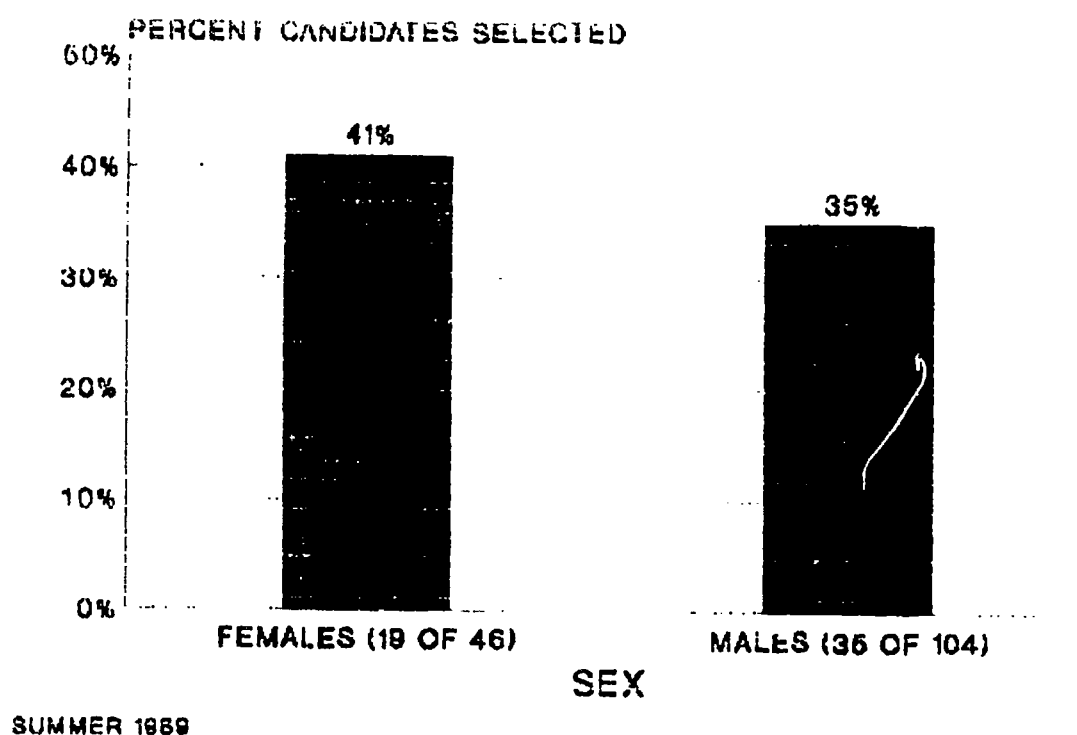
All raters and alternates were trained together using standardized rating materials and video tapes of sample performances by actors and others pretending to be applicants. Raters were given numerous opportunities to practice and compare their ratings. When all 11 raters seemed to be rating in substantial agreement, the final schedule of role plays was developed and the role play process began.

Role play administration lasted from the end of July, 1989 until the beginning of September of the same year. 148 applicants completed the role play process. After the final grading of the role play, an analysis of the results by rating panel group was made. The three rating panels were found to have very similar means and standard deviations. Therefore, no adjustment was made for rating panel variability.

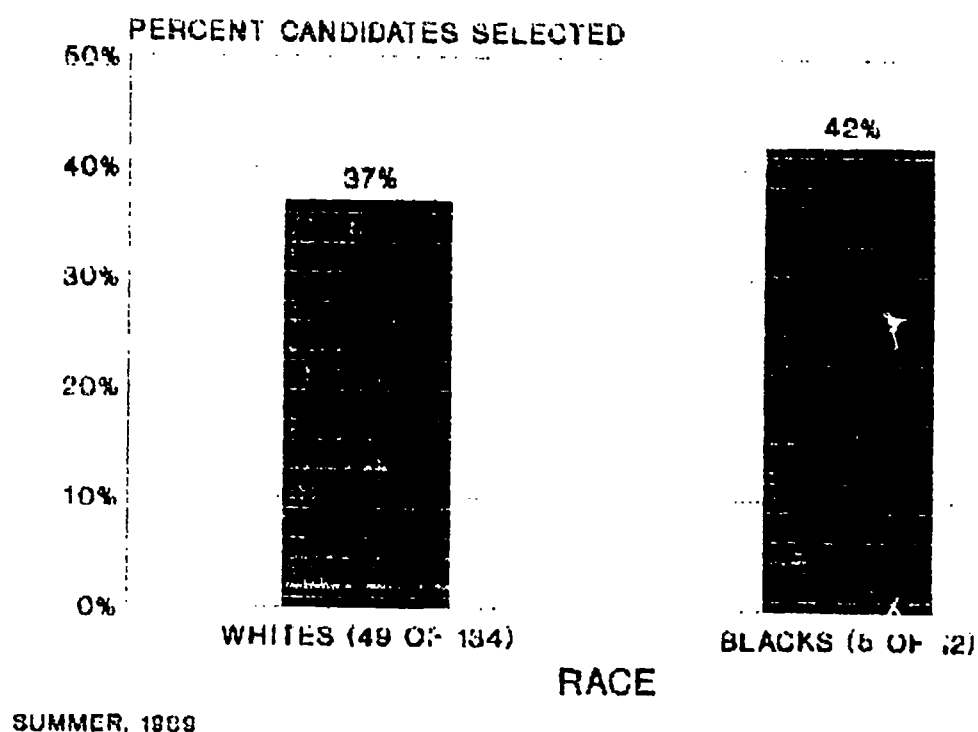
Prior to the final grading and ranking of the role play ratings subject matter experts were used to develop satisfactory and superior levels as potential cut points for the role play phase of the process.

During early September, 1989, the results of the role play were analyzed and compared against these two levels. It was decided to use the superior level cut point to determine which applicants would proceed to the final oral board interview phase of the process. 54 applicants were selected using this criteria. Figures 4 and 5 show the proportion selected for the oral board by sex and race respectively. As can be seen in those figures, both women and minorities were able to proceed to the oral board in representative numbers.

**FIGURE 4**  
**PERCENT SELECTED BY SEX - ROLE PLAYS**



**FIGURE 5**  
**PERCENT SELECTED BY RACE - ROLE PLAYS**



### Oral Board Interviews:

From the middle of September to early October, 1989, oral board interviews were conducted for the remaining 54 applicants. The interviews consisted of structured interview questions which were asked of all applicants. The questions used consisted of both situational problems and "behavioral interviewing" type questions which requested concrete examples from the applicant's past experience. Specific rating materials were developed for each question and raters trained in their use.

Raters for the interview consisted of upper level Supervisors and Chiefs from the Base Operations and Service Quality sections of the Transit Department.

Prior to the interviews, subject matter expert judgement was used to pre-determine a satisfactory level of performance for the oral interview.

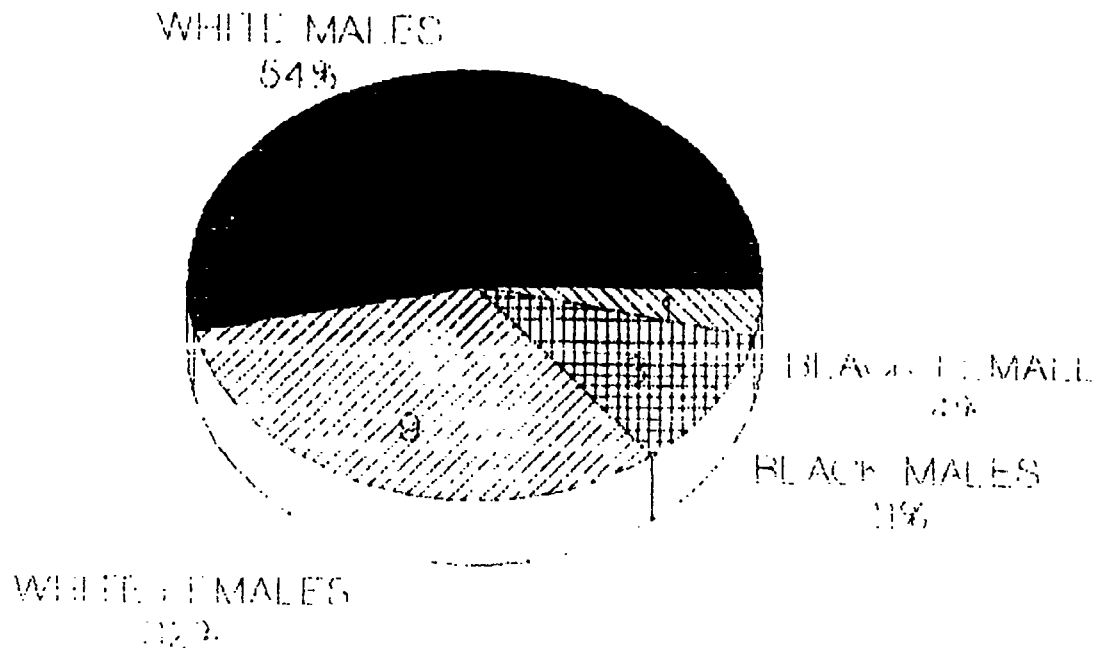
The interviews were completed during early October, 1989 and the results analyzed. It was found that all of the remaining applicants had passed the satisfactory level. The interview rating panel then decided to create a top down ranking of the remaining candidates based upon a composite of their performance during the written test, role play and oral board.

### Part VIII - Final Hiring Decisions & Creation of Hiring List:

The 54 individuals who participated in the oral board were then ranked in merit order based upon the composite scores mentioned above. Based upon budget projections, it was decided to immediately hire or place on the seniority list the top 28 ranked individuals. It was also decided to immediately hire 6 of those 28 individuals.

Figure 6 shows the composition of the final group of 28 selected for inclusion on the list. As can be seen from Figure 6, the final outcome of the process achieved a substantial increase in the number of women and a representative number of blacks.

**FIGURE 6**  
**FLIGIBLE LIST BY RACE AND SEX**



FALL 1989

## COMMENTS & CONCLUSIONS

### Transit Supervisor Selection - A Forewarning of Work Force 2000 Issues:

This most recent selection process for Transit Supervisors at Metro involved many of the issues which we all can expect to see more of due to the coming changes in the work force. For example, we had many internal applicants from a variety of race, sex, educational and age categories engaged in a highly competitive process for a relatively small number of promotional opportunities. At the same time, we were faced with significant affirmative action concerns and the constraints of tightly worded labor agreements. In dealing with these problems, Metro encountered issues that may well be part of dealing with Work Force 2000 selection procedures such as adverse impact in testing, when and how to set passing points, and whether or not to use race and/or sex conscious hiring procedures.

### Testing Methodologies & Adverse Impact:

Table 2 shows group means on the written, role play, oral interview and overall composite score portions of the selection process by race for blacks and whites (there were too few of other minority groups in the process to be meaningful).

TABLE 2  
MEANS & S.D.'s FOR DIFFERENT TESTING FORMATS BY RACE

Testing Component ----- Race	Written Test	Role Plays	Oral Board Interviews	Composite Scores
Blacks	N = 34-* X = 72.99 S.D. = 6.65	N = 12 X = 81.88 S.D. = 9.83	N = 5 X = 84.21 S.D. = 8.43	N = 5 X = 84.81 S.D. = 2.49
Whites	N = 150-* X = 81.29 S.D. = 5.95	N = 132 X = 79.80 S.D. = 10.17	N = 49 X = 79.57 S.D. = 10.12	N = 49 X = 83.37 S.D. = 4.50

$\bar{X}$  = Mean; N = Number of Cases; S.D. = Standard Deviation; \* =  $P < .05$

Like many forms of written testing, the written test portion of this process had an adverse impact on blacks. The role play and oral board portions of the process, however, did not have impact. If anything, some blacks performed well enough in these portions to make up for their written test scores, thereby scoring high enough to be hired or placed on the seniority list.

Table 3 shows Means and Standard Deviations on the various portions of the process by sex.

TABLE 3  
MEANS & S.D.'s FOR DIFFERENT TESTING FORMATS BY SEX

Testing Component ----- Sex	Written Test	Role Plays	Oral Board Interviews	Composite Scores
Females	N = 58 X = 80.80 S.D. = 7.37	N = 46-* X = 82.81 S.D. = 8.39	N = 19 X = 80.67 S.D. = 6.66	N = 19 X = 84.19 S.D. = 4.20
Males	N = 135 X = 79.23 S.D. = 6.69	N = 102-* X = 78.73 S.D. = 10.44	N = 35 X = 79.64 S.D. = 11.49	N = 35 X = 83.13 S.D. = 4.45

X = Mean; N = Number of Cases; S.D. = Standard Deviation; \* =  $P < .05$

As can be seen from Table 3, females as a group did as well as or better than males throughout all stages of this process. The only factor which tended to limit the number of women placed on the list in this process was the relatively low number of them who competed in the process. This recruitment did make significant progress from previous ones in recruiting women, however.



Table 4 shows a matrix of the correlations between candidate scores on the written test, role play, and oral board portions of the process.

TABLE 4

CORRELATION MATRIX - COMPONENTS OF THE PROCESS

Testing Component	Written Test	Role Plays	Oral Board
Written Test	1.00		
Role Plays	.174 *	1.00	
Oral Board	.098	.279 *	1.00

\* =  $P < .05$

These results indicate that there may be only a small correlation between the written test and the role play and a somewhat larger one between the role play and oral board. When using a multiple step selection process, it is important to know the degree of correlation between different steps. If the steps are relatively unrelated (i.e. - not correlated), then using one step to screen out candidates before they have a chance to compete in the remaining steps increases the likelihood of losing candidates whose overall performance is good enough for final selection.

Caution should be exercised in interpreting the above results, however, because both the role play and oral board portions of the process could be affected by range restriction due to the use of passing points in the preceding portion of the process. This would tend to give an inter-correlation coefficient lower than it actually is.

Methods for Achieving Affirmative Action:

During the last few years there seems to have been an increase in the legal scrutiny given to the use of race and sex conscious actions to achieve affirmative action goals in selection. The results of this process suggest that there are alternatives for at least some types of jobs.

In this process, role plays and oral boards seemed to give minority applicants more of an "even playing field" than written testing. Both role plays and oral boards, however, require considerable agency resources in staff and time. This can be a serious problem if large numbers of applicants are involved.

One way of compromising between the risks of adverse impact found in written testing and the resource requirements of role plays or oral boards can be found in the creative use of passing points. In this process, pre-determined "satisfactory" and "superior" pass points were used to tailor the number of applicants who would proceed to the next hurdle in the process while still maintaining reasonable and defensible standards of performance. This method allows enough flexibility to include sufficient numbers of protected groups applicants in the next phase when necessary yet still demonstrates a job related pass point which will provide defense against both discrimination and "reverse discrimination" complaints.

An important component of the success of this process revolved around the original recruiting efforts. In order to hire affirmatively with these types of methods, an agency must recruit relatively large numbers of protected group applicants. In this case, the agency made a major effort to recruit affirmatively, especially for women. Fortunately, these efforts paid off in dramatic increases in the number of women candidates and a corresponding increase in the probability of affirmative hiring.

### SUMMARY

The results of this process indicate that, if an agency is willing to invest significant resources in certain types of recruiting and selection procedures, affirmative action goals may be achieved without race or sex conscious actions. These methods also allow an agency to demonstrate that affirmative action was achieved without lowering selection standards for a particular group. In addition, they help avoid or defend against from both discrimination and "reverse discrimination" complaints while achieving affirmative action goals.



## ABSTRACT

This paper describes a job sample performance test developed to select individuals for the classification of Visually Handicapped Resource Assistant. Incumbents in the position spend the majority of their time reading for blind students enrolled in typing courses. The performance test consists of two self-taught study sessions and an evaluation test. During the study sessions, candidates teach themselves to act as the "eyes" of blind students by learning to read and completely describe all printed assignments for the pupils. Following the second session, candidates are required to read several paragraphs aloud for two evaluators. Final scores are determined by the candidates' ability to accurately and clearly read the text. Practical considerations of test administration are also discussed in the paper.

## A Work Sample Performance Test That Truly Recreates The Job

### The Classification:

The classification of Visually Handicapped Resource Assistant was created as part of the Los Angeles Unified School District's Visually Handicapped program; a program which operates at several of the District's occupational centers. The centers offer a variety of courses designed to help students learn and refine business skills. Typically, students enroll in typing, computer operations, English and math courses.

The primary goal of the Visually Handicapped program is to offer assistance to the visually impaired students who enroll and are mainstreamed into the occupational classes. It is particularly important that the handicapped students receive extra assistance and guidance with their classwork in order to perform well in the classes. The Visually Handicapped Resource Assistants provide such help by accompanying students to class, reading all class materials such as books and hand outs to the students, and reinforcing lessons learned in the class both on a one to one and group basis.

Probably the most challenging and critical responsibility for the Visually Handicapped Resource Assistants is aiding the visually impaired students in typing classes. All visually impaired students are strongly encouraged to enroll in typing and computer classes, yet, it is frustrating for blind students to learn how to set up the typewriter, use correct hand placement on the keyboard, and type text which they cannot see. The Visually Handicapped Resource Assistants must continually check that the students have properly set up the equipment and that their hands are placed correctly on the keyboard. Additionally, the Assistants must read all assignments from the typing textbook as the students type the lessons. In essence, the Assistants act as the typist's eyes, describing for the students every detail of the text and prompting the students when to type punctuation marks, capitalize words, return the carriage, insert a hyphen or parentheses, and so on.

Minimum qualifications for the position of Visually Handicapped Resource Assistant include one year of previous experience as a classroom aide, or a volunteer or student teacher in a school for the blind. Candidates must also have taken twelve college units in a number of courses related to the position; courses in English, psychology, special education, computer operations and typing would be acceptable. Candidates possessing a teaching credential or who have graduated from college with a major in psychology or sociology are also considered qualified.

#### History of the Examination:

Prior to 1989, the employment examination for the position of Visually Handicapped Resource Assistant consisted of a single test part, an interview. Candidates were asked a number of general questions relating to their background and job-related skills. The candidates were assessed on several factors including job preparation, communication skills, dependability, and sensitivity toward visually impaired individuals.

The Directors of the Visually Handicapped program were dissatisfied with the quality of candidates on the eligibility lists. The Directors described some of the candidates as having such poor communication skills, that the students were unable to understand them. Further, several of the candidates who were hired, were let go because they were unable to learn how to successfully read to the blind students even after a long training period. The Directors stressed a need for a more comprehensive exam which would assess the candidates oral communication skills more thoroughly and would measure the candidates ability to learn to read to the handicapped students.

#### Job Analysis:

In order to begin the development of the new employment examination for this position, a comprehensive job analysis was undertaken. Job analysis interviews were held with the Program Directors and with each of the incumbents. In addition, questionnaires were completed by the Directors and the incumbents, and job observations were carried out on several different occasions.

A list of thirteen knowledge, skills, and abilities were derived from the job analysis; the most critical of these KSAs included the ability to speak English clearly and distinctly using correct grammar and the ability to learn to read to blind students.

#### Development of the New Exam:

Based on the data and information collected during the job analysis, a new examination consisting of four test parts was developed. Candidates participating in the exam take a written test which includes sections on math, English usage, and knowledge of formats/rules used in typing. In addition, candidates take part in a job sample performance test which requires that they read several typed paragraphs as though they were reading a typing assignment to a blind pupil. Finally, the candidates participate in an interview and typing test.

#### The Job Sample Performance Test:

The job sample performance test requires the candidates to teach themselves how to read "timed writings" for blind students. A "timed writing" is a typing test administered to students in typing classes which helps the students gauge how quickly and accurately they can type. The students are given several paragraphs to type and they are to type as much of the text as they can in a five minute period. Visually impaired students of course require someone to read the "timed writing" to them.



The ability to read "timed writings" is not an ability required of Visually Handicapped Resource Assistants the first day on the job. In fact, there are few individuals who are trained to read to blind students, and it would be unreasonable to expect candidates to possess that ability. However, the Directors of the Visually Handicapped Program indicated that they had little time to train new employees thus they needed individuals who did possess the ability to learn to successfully read to the blind students. Thus, the job sample performance test was developed to test the candidate's ability to learn to read to the blind by first teaching the candidate how to read and then testing their reading skills.

Each candidate is mailed a study guide two weeks prior to the actual performance test. The study guide defines a "timed writing" and briefly describes the role of the Visually Handicapped Resource Assistant as a reader for the blind. The remainder of the guide outlines specifically how one should read a "timed writing". Examples include the following:

BEFORE READING THE ACTUAL TIMED WRITING YOU SHOULD:

- o read through and familiarize yourself with the text.
- o let the typist know if the text is single or double spaced.

DURING THE TIMED WRITING YOU SHOULD:

- o end each sentence by saying "period" or "question mark; the typist should automatically know to space two times and capitalize the first word in the next sentence.
- o say "RETURN" at the end of every line of type.
- o say "CAP" prior to reading a word that has been capitalized.
- o say "FIGURE" before reading a number.

In addition, a one paragraph example with a key is provided. The sample paragraph is relatively complex to read in that there are a number of prompts or commands that the candidates must remember to say to the student during the "timed writing". The key indicates each command and spells out exactly what should be said.

On the day of the performance test, candidates are given a practice period during which time they are encouraged to review the study guide. To supplement the guide, a study tape is provided. The tape is a recording of the sample "timed writing". Candidates are encouraged to read along with the tape as practice.

After the 45 minute practice period, each candidate is invited into the performance test, one at a time. Candidates are directed to read a relatively complex, four paragraph "timed writing" as though they were reading to a blind student. The text contains a number of hyphenated words, quotation marks and other characters, numbers, capitalized words, and compound words. The level of complexity of the text is typical of an actual "timed writing" which would be read on the job. There are a total of 66 prompts that the candidates should indicate while reading. Candidates are assessed on their ability to read a "timed writing" and their ability to speak English clearly and understandably.

While candidates are reading, two evaluators keep track of each prompt that the candidates accurately indicate. At the completion of each test, the evaluators compare scores to insure that they have counted accurately. All of the performance tests are tape recorded. If the two raters find a discrepancy in their scoring, they are instructed to listen to the tapes again to determine the correct score. This score is the final score for the factor of "ability to read a timed writing". The evaluators also assess the candidates' oral communication skills by assigning a score of excellent, strong, acceptable or weak in this area. Specific behavioral anchors are provided as descriptions of the excellent, strong, acceptable and weak candidate. Candidates must receive a score of at least 33 on the first factor and an acceptable on the second factor to pass the performance test.

All evaluators attend a rater training session prior to rating the performance test. Although the evaluators are incumbents in this position and therefore are familiar with how to read a "timed writing", they are not familiar with the actual testing process. The training session is set up to allow the evaluators ample time to acquaint themselves with the test. The evaluators also practice scoring the test by listening to several pre-recorded examples of candidates reading the "timed writing". After each reading, the tape is stopped and the raters are asked to score the sessions as though it were an actual candidate competing in the exam. Scores are compared to insure that the scoring procedures are being followed.

#### Discussion:

As with any employment examination, the job sample performance test for the position of Visually Handicapped Resource Assistant has its strengths and weaknesses. The test seems to be a better method of assessing candidate's abilities to read timed writings and their communication skills than the interview which was previously used. In fact, the Directors of the Visually Handicapped Program have expressed great satisfaction with the new test. Furthermore, the Directors have indicated that they have not had to train new incumbents to read for visually impaired students as extensively as they have had to in the past because the candidates have trained themselves prior to the test.

Additionally, the study guide may act as a realistic job preview for the candidates. Candidates who are not familiar with the classification and find that they are not interested in reading to blind students, tend to self-screen themselves out of the selection process. Finally, the test is easy to set up and administer; aside from the practice period, the test only takes about ten minutes.

On the other hand, there are several possible limitations which have arisen relating to the test. First of all, a number of candidates originally qualified to take the exam, did not show up to participate in the test. It may be that the performance test is somewhat intimidating to the candidates. Also, there is no way to control the amount of time candidates devote to reading the study guide. Mailing the guide two weeks prior to the test may not give the candidates enough time to review the material. Further, there is no statistical evidence that the passpoint of 33 in the area of "ability to read a timed writing" is appropriate. Finally, there are a limited number of individuals who are qualified to evaluate the candidates. If there is a large candidate population, this may cause some problems.

The performance test as well as the rest of the examination for the position of Visually Handicapped Resource Assistant has recently been developed and only tried out on a handful of candidates. The School District is anxious to use the test again in order to find out more about its validity and usefulness. To achieve this goal, studies should be conducted to determine if the present passpoint should be adjusted. The test might be more successful if study guides are sent out earlier to allow candidates more time to learn the material. Candidates may also be invited to workshops or job-orientation meetings set up to provide more information about the classification and the testing procedure. Hopefully, the meeting will result in a higher number of candidates participating in the testing process .

**LIST OF DUTIES/RESPONSIBILITIES/  
VISUALLY HANDICAPPED RESOURCE ASSISTANT**

- |   |     |
|---|-----|
| 1. Reads classroom materials such as textbooks and hand-outs to visually impaired students so that students may keep up with the other individuals in the class.  | 45% |
| 2. Tutors students on a one-to-one basis in basic educational areas and in the use of business machines such as typewriters and computers.  | 15% |
| 3. Tutors groups of students in basic educational areas and in the use of business machines such as typewriters and computers.  | 10% |
| 4. Copies, enlarges, tapes, and types classroom and tutoring materials so that the visually impaired students will have access to the printed materials used in class.  | 10% |
| 5. Acts as a liaison between teachers and visually handicapped students in the classroom to insure that there is good communication between the student and the teacher.  | 5%  |
| 6. Assists a resource teacher in developing a variety of instructional materials by researching, extracting or rephrasing portions of textbooks and classroom instructional sources.                              | 5%  |
| 7. Investigates problems that are being encountered by aides, students and classroom teachers, and recommends solutions to the Program Coordinator.   | 5%  |
| 8. Carries out other duties as assigned such as providing administrative support to the program coordinator, driving visually impaired students to job interviews, and walking with students to and from classes. | 5%  |

LIST OF ESSENTIAL KNOWLEDGE, SKILLS AND ABILITIES FOR THE  
POSITION OF VISUALLY HANDICAPPED RESOURCE ASSISTANT

- o Ability to speak English clearly and distinctly using correct grammar
- o Ability to read "timed writings" and other text to blind students
- o Ability to work independently
- o Knowledge of basic English grammar, spelling, and punctuation
- o Knowledge of basic math
- o Ability to maintain an adequate attendance/punctuality record
- o Ability to interact effectively with visually handicapped students, teachers, administrators, and others
- o Knowledge of the operation of typewriters
- o Knowledge of formats, punctuation, spacing, and other rules used in basic typing
- o Ability to type accurately
- o Ability to make quick and rational decisions on the job
- o Knowledge of the operation of computers
- o Ability to teach others

\*\*Please note: KSAs are listed in order of criticality.

PREVIOUS VERSUS CURRENT EXAMINATION FOR THE POSITION OF VISUALLY  
HANDICAPPED RESOURCE ASSISTANT

PREVIOUS EXAM	CURRENT EXAM
<ul style="list-style-type: none"> <li>o Interview               <ul style="list-style-type: none"> <li>KSAs tested for:</li> <li>- Job Preparation</li> <li>- Sensitivity to the needs of Visually Impaired</li> <li>- Oral Communication Skills</li> <li>- Dependability</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>o Written Test               <ul style="list-style-type: none"> <li>KSAs tested for:</li> <li>- Knowledge of Basic Math</li> <li>- Knowledge of English Usage</li> <li>- Knowledge of formats, punctuation, spacing and rules used in basic typing</li> </ul> </li> <li>o Job Sample Performance Test               <ul style="list-style-type: none"> <li>KSAs tested for:</li> <li>- Ability to Read to Blind Students</li> <li>- Oral Communication Skills</li> </ul> </li> <li>o Interview               <ul style="list-style-type: none"> <li>KSAs tested for:</li> <li>- Ability to Work Independently</li> <li>- Ability to Make Quick and Rational Decisions</li> <li>- Ability to Interact Effectively With Others</li> <li>- Ability to Maintain a Good Attendance and Punctuality Record</li> </ul> </li> <li>o Typing Test               <ul style="list-style-type: none"> <li>KSAs tested for:</li> <li>- Ability to Type Accurately</li> </ul> </li> </ul>



## NEW DEVELOPMENTS IN PERSONALITY MEASUREMENT

Robert Hogan

University of Tulsa  
(918-584-5992)

IPMAAC--San Diego--June, 1990

When HR professionals think of personality psychology, they tend to think of psychoanalysis and names like Freud, Jung, and Maslow. When they think of personality measurement they think of measures like the Rorschach, the TAT, and the MMPI. And when they think of the relevancy of all of this for the world of work, they draw a blank--or they dismiss the entire business as irrelevant.

There is some wisdom to this conclusion. Traditional personality psychology was designed to explain how people fall apart and to provide clues for putting them back together. Traditional personality psychology is not particularly relevant for understanding the goals, values, motives, and aspirations of normal people, including working adults. Specifically, psychoanalysis assumes that the most important generalization we can make about people is that everyone is neurotic, and the most important problem in life is to overcome our neuroses. As for traditional personality assessment of the MMPI variety, years of research indicate that measures of psychopathology are poor predictors of most aspects of job performance.

Consequently, there is a certain scepticism in the HR community regarding the utility of personality measurement for assessing or predicting components of job performance. But this scepticism concerns a particular definition of personality and a particular orientation to measurement. There are, however, alternatives to these definitions that lead to different conclusions.

As one alternative to psychoanalysis consider the point of view called interpersonal theory. This viewpoint argues that people always live in groups, that every group has a status hierarchy, and that, therefore, the major problems in life concern getting along with others and getting ahead--i.e., achieving status and social acceptance.

Concerning the relationship between personality and job performance, there is a major confusion in most people's minds about what personality is. It is actually two things. On the one hand, personality refers to a person's reputation (e.g., Charles has a colorful personality, Fred is a bit depressed). Personality in this sense is expressed in trait words. These trait words reflect how the person has behaved in the past and how he or she is likely to behave in the future. Reputations provide a basis for forecasting future behavior, including job performance. On the other hand, the word personality refers to whatever it is inside people that causes them to behave the way that they do.

I believe that the first definition of personality--as reputation--is something we can study scientifically. I am not sure we can study the second in a rigorous way.

But with regard to personality as reputation, there is some interesting news to report. Factor analytic research over the past 15 or 20 years has converged on the notion that there is a consistent underlying structure to reputation. And the structure is a cultural universal--the structure of reputation is the same whether we study it in Japan, Korea, the African outback, or among the Inuit. Virtually every personality researcher in the world agrees that the structure of reputation can be represented in terms of five (or six) broad dimensions, which are described at the top of Figure 1. All of these dimensions are associated with popularity or success or both, and we, therefore, refer to these dimensions as the "bright side" of personality.

There are four points to be noted about the relationship between these "bright side" dimensions and job performance:

1. If you want to study the relationship between personality and on-the-job performance, you must consider all 5 (or 6) dimensions. It won't do to use as a measure of personality a self-esteem scale, or an honesty scale, or a special purpose scale such as locus of control or self-monitoring. You need to use all 5 (or 6) dimensions. This is an important rule that is violated frequently, and when violated often results in the researcher concluding that "personality" is unrelated to job performance.
2. The five dimensions of personality are relevant to different criterion data--see Figure 2. The relevant attribution for Figure 2 is an Army technical report by Leaetta Hough at PDRI in Minneapolis. Because performance criteria in organizations vary widely, different aspects of personality are relevant to different criteria.
3. When you do the research correctly, then selection procedures based on these measures yield a reasonable pay off in terms of enhanced productivity and reduced overhead. Figure 3 contains two examples. The top part of the Figure shows the productivity levels of two groups of insurance claims examiners, one group having been chosen in the typical way--with an interview--and the second having been chosen with a personality measure.
4. Finally, there are a number of jobs in the world in which it is hard to know what counts as good performance, but it is easy to know what counts as poor performance. On jobs like insurance claims examiner (or school bus driver) (or air traffic controller) the most salient dimensions or aspects of performance are the errors--so that it is hard for the organization to specify good performance but easy to detect poor performance. This points to the relevance of the "dark side"--see the bottom of Figure 1. "Bright side" measures can also be used to detect poor performers; the bottom of Figure 3 provides an example of how a "bright side" measure can reduce bogus worker compensation claims. The bottom of Figure 3 is quite instructive. Idaho has a depressed economy.

The state uses an IQ test (the GATB) to screen applicants. The job at the state hospital for the profoundly retarded was unpleasant. Relying almost exclusively on an IQ test for pre-employment screening, the state hired a number of bright but marginally delinquent people into the job; these people didn't like the job and found ways to aggress against the system.

Let me summarize what I have said thus far. First, personality should be conceptualized in terms of the reputational factors that are associated with status and social acceptance in the social and occupational groups in which one takes part. Second, when personality is conceptualized in this way and psychometrically adequate measures of these factors are developed, the measures will predict those positive aspects of job performance that are relevant to the constructs. These first two points depart from the conventional wisdom of the HR "community." Third, these "bright side" measures are less useful for predicting negative aspects of job performance.

This brings me to my last general point. That is that, within a population of people who interview well and look great on measures of normal personality, there will be a subset of people with substantial "dark side" problems--which are exceedingly difficult to detect in an interview or with "bright side" measures. Consider the following example. Richard Berendzen was president of American University (in Washington, D.C.) for 10 years--he just stepped down. He is a hard working, charming, charismatic man who raised entering student test scores, paid for and built 8 new buildings, quadrupled the university endowment, and generally enhanced the visibility of a once pretty mediocre university. By all accounts, Berendzen was one of the premiere university presidents in the country. He resigned on April 10th because the Fairfax County (Virginia) police had identified him as the man who had made a series of obscene phone calls, some of which involved sexual fantasies about children. Quoting the May 16th Chronicle of Higher Education, "How...could a man lead such disparate private and professional lives? How could a man who so eloquently outlined his vision for a global university one moment engage in the kind of behavior being talked about the next, without providing a single clue that something was wrong....'There is a professional man and a private man. The private man has a problem and the professional man has a great record'"said the Chairman of the University's Board of Trustees.

We have been studying this dark side issue for the last two years and we have some preliminary results to report. First, we have found it useful to think in terms of what are known as the DSM III, Axis 2 personality disorders. A personality disorder is an elaboration of a normal personality characteristic. These are not neuroses, they are quirks and idiosyncrasies. I believe most people have personality disorders--shyness, stubbornness, defensiveness, problems with authority, social insensitivity. These quirks are difficult to detect with interviews, and they are largely unrelated to scores on either the MMPI or well constructed measures of the bright side.

The bottom part of Figure 1 describes the most common "dark side" dimensions. To give you a feeling for how these work, consider two

examples from some recent consulting we have been doing. Figure 4 is the HPI profile of the R & D manager of a high tech firm in the east. He has a Ph. D. in physics from a famous university; he is exceptionally well-trained and well-respected for his technical expertise; and he is the best-liked manager in his organization. He is also a "high likeability floater". These are people who specialize in good relations rather than productivity. His boss complained that he is always late with projects, his group is underproductive, he is reluctant to take on new projects on the grounds that the company is compromising the quality of his lab's work. A glance at his dark side profile (Figure 5) shows elevated scores for Passive Aggressive and Perfectionism. Such people are obsessive-compulsive procrastinators.

A second example is a man who is a world-class salesman. A glance at Figure 6 shows him to be colorful, ambitious, but hard nosed and insensitive. This person just derailed as a manager of marketing because his staff was in open revolt. Figure 7 shows why. The organization lost a great deal of money before this narcissistic manager derailed.

What are the new developments in personality measurement? There are essentially two, and they are important. The first concerns the emergence of the Five Factor model as a way of structuring concepts and measures of normal personality. All existing measures of personality assess the same five dimensions with more or less relative efficiency. These five factors are differentially related to job performance, depending on what kind of criterion data are chosen. And it is a sign of being out of touch if you define and assess personality with something less than at least these 5 dimensions.

Second, there is a domain of personality that is independent of that assessed by measures of normal personality and by the MMPI. This is the domain of the personality disorders. Unlike the MMPI and traditional measures of psychopathology, well-constructed measures of personality disorders provide information that in fact is relevant for job performance.

And for you keen observers who believe you have detected an inconsistency in my argument yes, the personality disorders can be reduced to the five factor model, but the manner in which that is done requires a long discussion that is best saved for another time.

Consider the following real life scenarios:

(a) A truck driver with a fine record and good credentials stops his rig in the middle of the New Jersey Turnpike, leaps out with a loaded pistol muttering imprecations against management and fends off curious people who wonder why this truck is stopped in the middle of the highway.

(b) A pleasant and upwardly mobile young executive kept notes on his boss and sent the notes secretly to his boss's supervisor.

(c) A bright and well-qualified secretary is also a religious fanatic; she severely disapproves of the behavior of her office mates and her boss and complains to everyone she knows about their behavior.

(d) A well-qualified young accountant secretly sells the client list and pricing information of his company to its primary competitor

Perhaps the most interesting new development in personality measurement is the ability to detect these tendencies in advance (the truck driver is an example of a paranoid personality, the religious fanatic is a schizotypal personality, and the two junior executives who betray their bosses and companies are a combination of passive-aggressive and narcissistic personalities). The link between personality disorders and job performance is one of the most interesting lines of new research in industrial psychology.



## PERSONALITY AND OCCUPATIONAL PERFORMANCE

Screening for Competence: The Bright Side

<u>HPI Scale</u>	<u>Descriptors</u>
Intellectance	Concrete minded vs. Curious
Prudence	Impulsive vs. Meticulous
Ambition	Self-satisfied vs. Status Seeking
Sociability	Socially Reticent vs. Extraverted
Likeability	Hard-nosed vs. Diplomatic
Adjustment	Anxious vs. Self-Confident

Screening for Potential Problems: The Dark Side

<u>PROFILE Scale</u>	<u>Typical Problem</u>
Interpersonal Insensitivity	Politically Obtuse
Argumentative	Vengeful
Unstable Relationships	Door Slammer
No Common Sense	Bad Judgement
Attention Seeking	Noisy Fan
Arrogance	Ignores Feedback
Fear of Failure	Indecisiveness
Dependency	Requires External Support
Passive-Aggression	Procrastination
Perfectionism	Unable to Prioritize
Untrustworthiness	Delinquency

Source: Hogan, J., Hogan, R., & Arneson, S. Test Validity  
Yearbook (in press). F. Landy (Ed.). Erlbaum



# Performance Implications of the Five Factors of Personality

<u>Dimension</u>	<u>Criteria</u>
Intellectance	Training and/or Academic Performance
Prudence	Disciplinary Problems/Honesty
Ambition/ Sociability	Leadership/Upward Mobility
Likeability	Popularity/Peer Acceptance
Adjustment	Supervisor's Ratings

# IMPACT OF VALID PERSONALITY-BASED SELECTION SYSTEMS

---



---

## INSURANCE CLAIMS EXAMINERS

AVERAGE TRAINING SCORE		PER CENT CLAIMS PROCESSED		
		3 weeks	6 weeks	9 weeks
Non-tested	672	40%	46%	54%
Tested	684	66%	83%	97%

Utility per new hire per year = \$105,000

20 hired 1989 = \$2,100,000

---

## STATE OF IDAHO - INSURANCE COMMISSION

Claims Filed	Average Cost/Claim	Number of incidents	Total
High Scorers .8	\$710	59	\$41,890
Low Scorers 1.5	\$1037	144	\$149,328

Savings for 7% of workforce = \$107,438

Savings for 93% of workforce = \$1,427,390

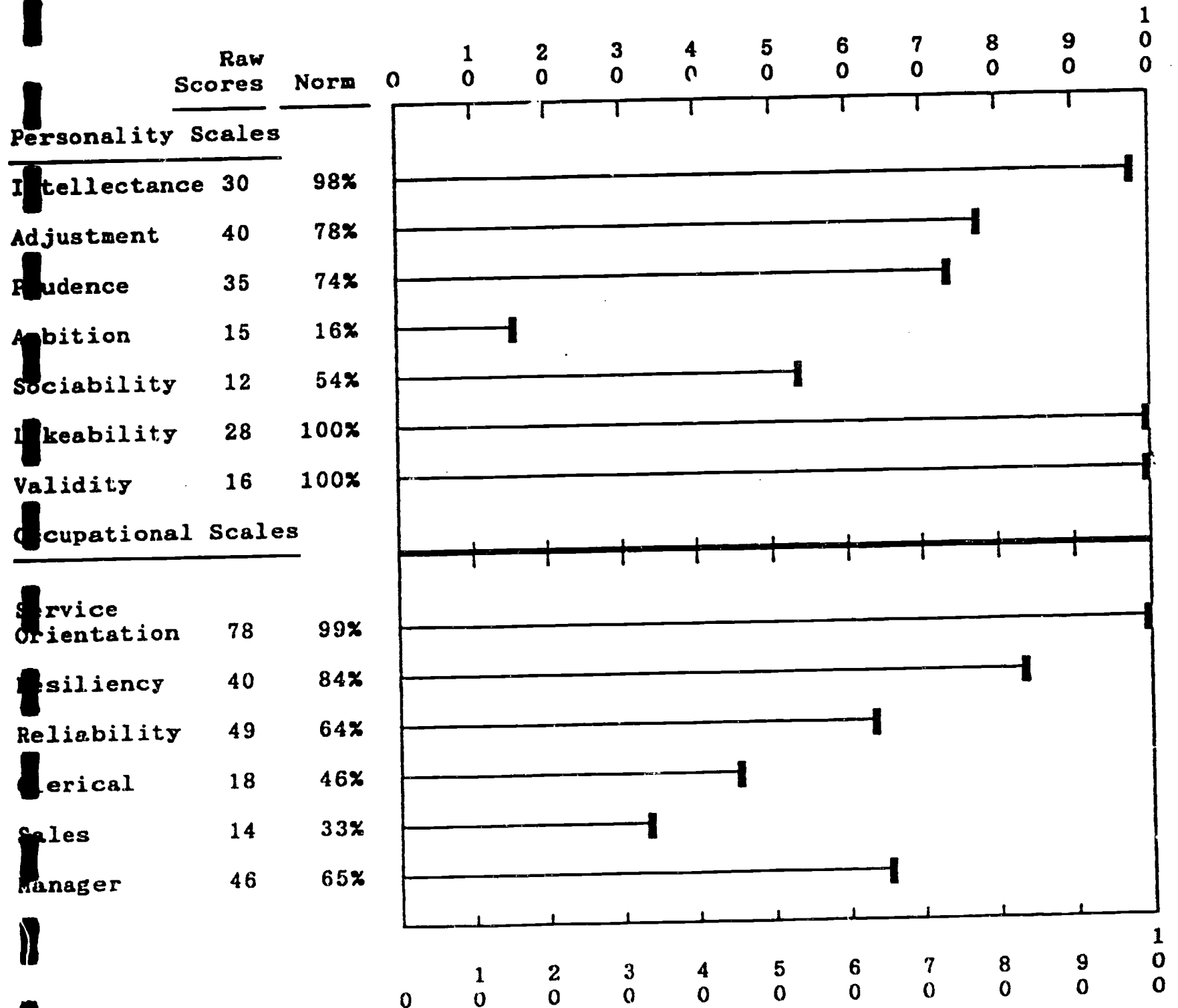
Total Savings = \$1,534,828

# HOGAN PERSONALITY INVENTORY

## GRAPHIC PROFILE

Name: [REDACTED] High Likeability Floater  
 Date: 31 May 1990

Page: 1



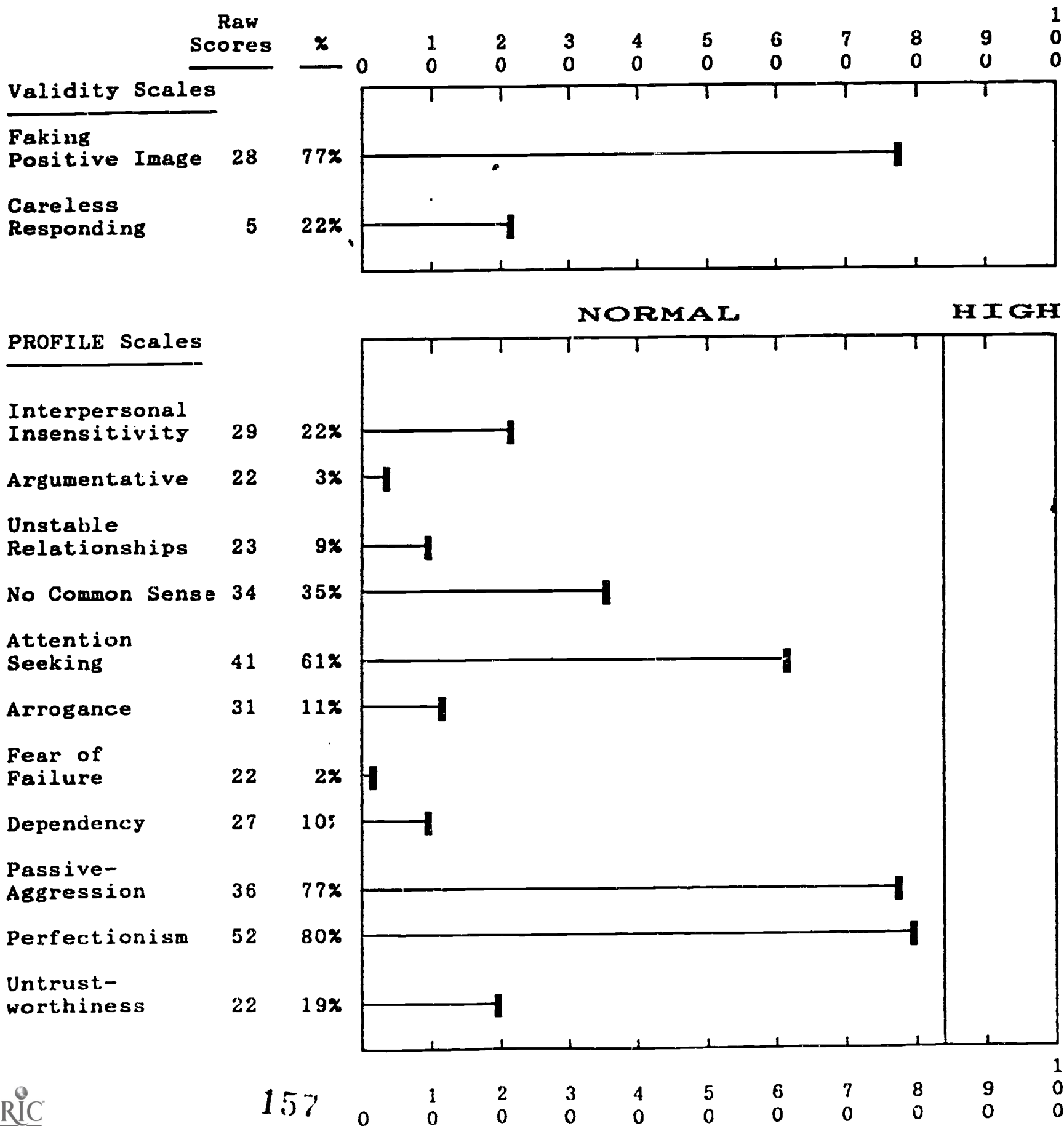
# FIGURE FIVE P R O F I L E

## OCCUPATIONAL REPORT

### GRAPHIC PROFILE

Name: XXXXXXXXXX High Likeability Floater  
Date: 31 May 1990

Page: 1

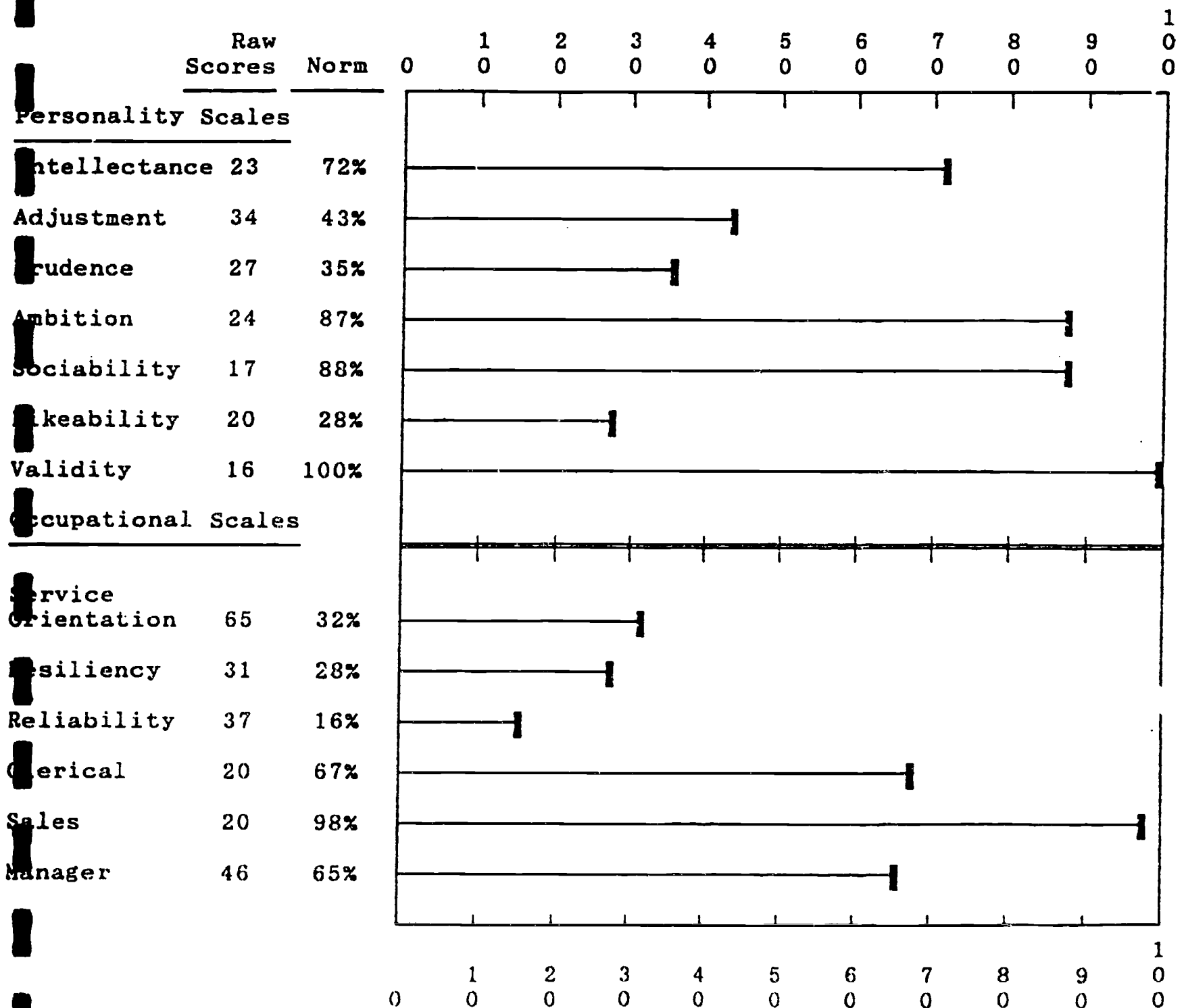


## HOGAN PERSONALITY INVENTORY

## GRAPHIC PROFILE

Name: [REDACTED] Narcissist  
 Date: 1 June 1990

Page: 1



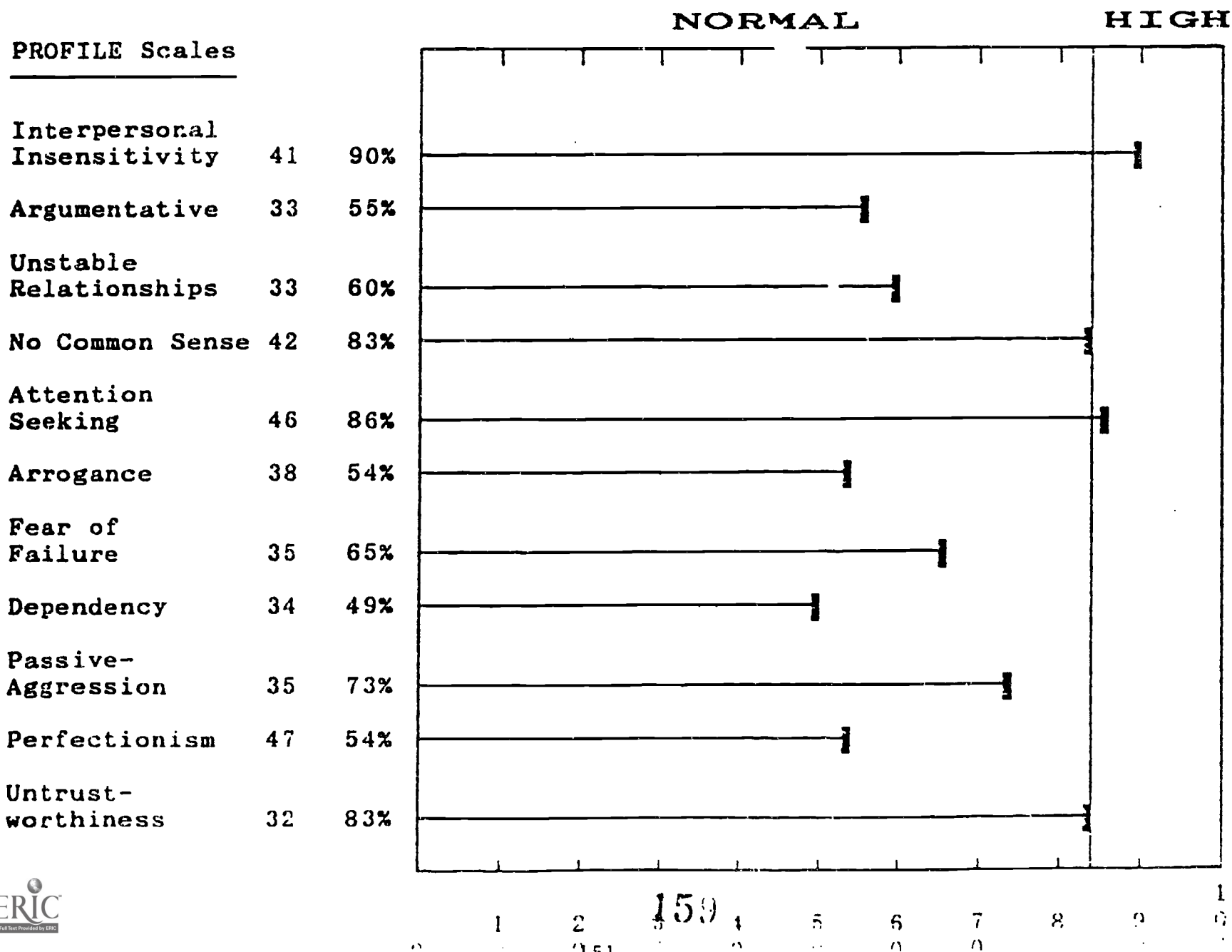
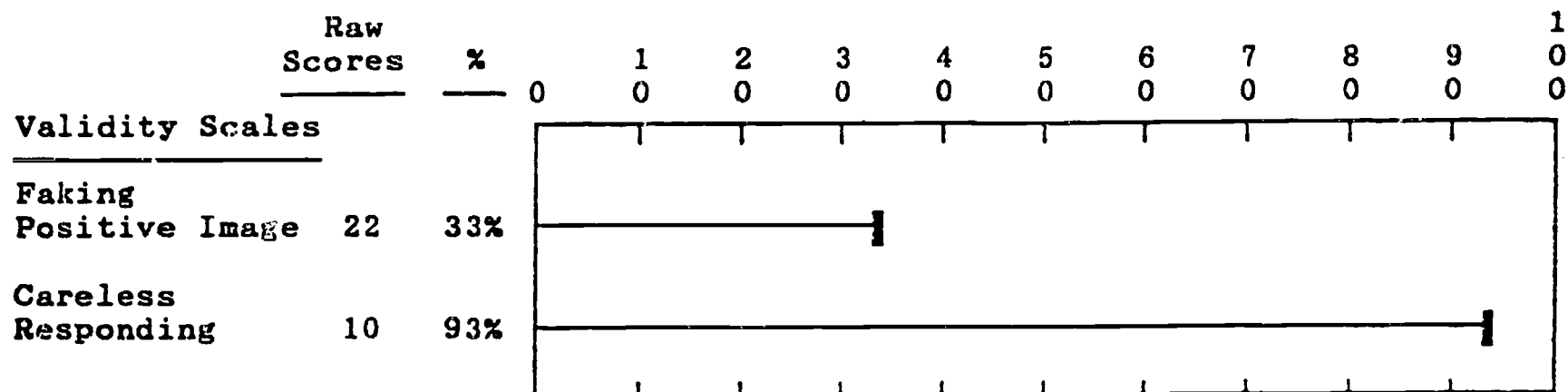
## P R O F I L E

## OCCUPATIONAL REPORT

## GRAPHIC PROFILE

Name: [REDACTED] Narcissist  
 Date: 1 June 1990

Page: 1





The Prospective Employee Potential Inventory:  
a Validation Study with School Bus Drivers

Thung-Rung Lin, Teresa F. Doyle and J. Mark Howard  
Los Angeles Unified School District

Abstract

A summary of the single largest validation study of the Prospective Employee Potential Inventory (PEPI) is presented. Data were collected from 649 school bus drivers employed by a large, urban school district in the western U.S. Hard and soft performance data obtained from the bus drivers' personnel files were correlated with their scores on the PEPI to investigate the PEPI's utility as a screening tool in the selection of future drivers. Results did not demonstrate a sufficiently strong relationship between predictors and criteria to justify including the PEPI in the selection process.

As one of its region's largest employers of bus drivers, a large, urban school district has continually sought to improve its selection practices for school bus drivers in an effort to promote safety and reliability, while producing eligible lists with sufficient numbers to satisfy the needs of the district. This district hires nearly 200 new bus drivers annually and the present multiple hurdles examination, consisting of a written knowledge test, two-part performance test (bus inspection and driving), and structured interview, has been effective in selecting individuals who are proficient in driving school buses. However, it has become evident in recent years that a lack of driving ability is not the only characteristic that causes incumbents to perform badly as school bus drivers. In fact, poor driving is not even a major contributor to bus driver failure on the job, as management in the district's Transportation Department has complained that bad bus drivers typically distinguish themselves by being frequently late or absent, or having poor working relationships with other school personnel and the public. These and similar factors have been suggested elsewhere as conducive to failure among bus drivers employed in municipal transit authorities (Ash, et al., 1988).

The results of a job analysis of the Bus Driver classifications in this school district support these observations. Recently, more than 1,000 incumbent bus drivers completed a job analysis questionnaire which, in part, asked them to rate knowledge, skills, abilities, and personal characteristics (KSAPs) associated with their jobs in terms of relative importance to performance. A four-point scale, ranging from "not important" to "exceptionally important," was used for the ratings. The 109 KSAPs included in the questionnaire were then roused into clusters yielding nine dimensions: technical job knowledge, technical proficiency, communication, human relations, detail orientation, flexibility, decisiveness, dependability and trainability. As can be seen in Table 1, on the average, the dimensions of decisiveness, dependability, flexibility, human relations, and trainability were rated higher than technical job knowledge for relation to performance. Dependability, decisiveness, and trainability were also rated as high or higher than technical proficiency, the KSAPs associated with the actual operation of a school bus.

Table 1. Average dimension ratings for relation to performance as a Bus Driver.

Dimension	Mean	SD	N
Dependability	3.29	.72	929
Decisiveness	3.11	.70	1010
Trainability	3.11	.76	969
Technical Proficiency	3.11	.65	1015
Flexibility	3.08	.73	992
Human Relations	3.07	.71	985
Technical Job Knowledge	2.99	.71	1013
Communication	2.80	.89	980
Detail Orientation	2.74	.85	1000

Note: Ratings were assigned as (1) not important, (2) slightly important, (3) important, (4) exceptionally important.

Dimensions are defined as follows:

Dependability - dedication to optimal work performance at all times.

Decisiveness - readiness to make decisions, take action, and commit oneself.

Trainability - understands and puts into practice continuing instruction, guidance, or direction.

Technical Proficiency - knowledge, skills, abilities, and personal characteristics necessary to safely operate a school bus and complete a route.

Flexibility - effectiveness in the management of stress accompanying the performance of job duties.

Human Relations - effectiveness in establishing and maintaining efficient, positive working relationships with supervisors, students, school personnel, parents, the public, and co-workers.

Technical Job Knowledge - familiarity with and understanding of the components, maintenance requirements, and capabilities of different types of buses.

Communication - effective expression in writing, and orally in both individual and group situations.

Detail Orientation - attention to routine or unspecialized, usually non-technical duties, such as personnel requirements, paperwork, and bus cleanliness.

Though the structured interview portion of the present examination does address attendance practices and interpersonal skills, it was hoped that a less conspicuous measure might prove useful in detecting candidates who are undesirable in these respects. It was also hoped that such a measure could provide an indication of a candidate's susceptibility to stress, or responses to stressful situations as, in job analysis interviews, incumbents cited stress on the job as the root of poor performance. Thus, it was suggested that a non-cognitive measure might be an appropriate addition to the selection process.

The use of non-cognitive assessments, such as personality tests, in the employment setting has a long and highly controversial history, as low to modest validities have generally been reported for these types of instruments (Schmitt, et al., 1984; Ghiselli, 1973). For example, Ghiselli (1973) reported a correlation of .26 using personality traits to predict job proficiency in a group of less than 499 Vehicle Operators. More recently, Schmitt, et al. (1984), in a meta-analysis of 62 validity studies with a total N of more than 23,000, reported an average validity of .15 for personality tests in the prediction of a variety of criteria such as performance ratings and turnover.

One criticism of the traditional personality tests and their use in the context of personnel selection is that their foundations in clinical psychology make them useful for the detection of pathological or abnormal behavior and less so in making accurate predictions about "normal" job applicants (Muchinsky, 1986). However, Hogan & Hogan and their associates (Hogan, Hogan & Busch, 1984; Hogan & Hogan, 1986; Hogan & Hogan, 1989a; Hogan & Hogan, 1989b) have developed personality-based measures that are derived from non-pathological human behavior and which may enhance the utility of non-cognitive assessments in the screening of candidates for employment.

The Prospective Employee Potential Inventory (PEPI) is a brief, self-administered inventory derived from the Hogan Personality Inventory (HPI). The PEPI consists of four scales titled Reliability, Service Orientation, Stress Tolerance, and Validity, which are intended to provide assessments of a candidate's maturity and conscientiousness, helpfulness and adjustment, adaptability, and consistency of responding, respectively (Hogan & Hogan, 1989b). The PEPI's subscales had been developed using incumbents from a number of different occupations, including hospital personnel, clerical workers, and truck drivers (Hogan & Hogan, 1986). Selected results from the truck driver studies indicated that individuals scoring high on the Service Orientation, Reliability, and Stress Tolerance scales typically received more commendations for work behavior than those scoring low (Hogan & Hogan, 1986). In addition, low scorers on Reliability were discharged from their company at a greater rate than high scorers (Hogan & Hogan, 1989a).

It was decided that the utility of the PEPI for the district's selection of bus drivers would be investigated. Specifically, a concurrent validity study was conducted using bus drivers' responses to the PEPI and performance criteria obtained from their personnel files.

### Method

Sample. Participants were 649 school bus drivers employed by the school district. They were paid for one hour at their usual hourly rates to complete the PEPI. All were self-selected volunteers from among the 1,038 who took part in task and KSAP analyses of the school bus driver classifications. Of the 531 who reported their genders, 50.7% were female. They ranged in age from 18 years to over 55 and, of the 519 who reported age, 78.5% were between the ages of 18 and 45. Of the 531 reporting ethnicity, 53.7% were African American, 22% were White, 20.5% were Hispanic, and 3.8% listed "other." Only 286 persons reported level of education, and 89.5% of these had at least completed high school. The average length of tenure with the Transportation Department (423 reporting) was 6.5 years.

These participants completed the PEPI in small, supervised groups ranging in size from five to 20.

Predictors. The Prospective Employee Potential Inventory is a 198-item, self-administered instrument published and marketed in 1989 by National Computer

Systems, Inc. as a tool for use in the selection of entry level employees. The PEPI is derived from a recombination of homogeneous item clusters (HICs) originally contained in the Hogan Personality Inventory. The PEPI is not divided into the same empirical scales as the HPI, but all items contained in the PEPI were initially part of the HPI. Each item consists of a single first-person statement, and respondents are asked to report whether they agree or disagree with the statement by answering true (agree) or false (disagree). Responses are aggregated to produce scale scores for Service Orientation (SO), Reliability (RE), Stress Tolerance (ST), their associated HICs, and for the Validity scale.

In personal communications with an author of the PEPI, it was suggested that its psychometric properties could be improved using revised (unpublished) scales for SO, RE, and ST, as well as revised HICs (Hogan, 1990). Thus, in addition to the scores for SO, RE, and ST, scores for the following HICs were obtained: SO-Positive Affect (PA), SO-Unlikely Virtues (UV), SO-Sensitivity (SE), RE-Avoids Trouble (AT), RE-No Hostility (NH), RE-Attachment (AA), ST-No Somatic Complaint (NS), ST-Calmness (CA), ST-No Guilt (NG), ST-Not Depressed (ND).

Criteria. The performance measures utilized for this report were absences (AL), supervisors' performance evaluations (PE), self-reported traffic incidents (SRTI), positive performance recognition (PPR), negative performance recognition (NPR), hard criteria associated with driving performance (HDP), and soft criteria associated with driving performance (SDP). All were obtained from the participants' personnel files as part of the Bus Driver job analysis project. All variables but SRTI were controlled for the number of years the bus driver had been employed by the district. Data were collected from the school fiscal years 1981 through 1988 or, if the driver had been employed with the Transportation Department less than eight years, the entire length of tenure.

AL consists of the number of days a driver had been recorded as absent in the absence log. PE is the average overall performance evaluation, ranging from one ("below standards") to three ("exceeds standards"), assigned to the bus drivers by their immediate supervisors in annual reviews. Each driver is asked to voluntarily report any traffic citations and accidents they have incurred up to the time of application with the Transportation Department; the count of these incidents constitutes SRTI. PPR is a measure of a bus driver's total recognition for positive performance, including, for example, letters of commendation and departmental notices of outstanding performance. NPR is a measure of negative recognition, including letters of complaint and departmental notices of unsatisfactory conduct, for example. The components of both PPR and NPR were differentially weighted according to the relative impact each has on the bus driver's standing, and summed to arrive at a final value. HDP is a count of letters, warnings, license suspensions and revocations (differentially weighted) from the State Department of Motor Vehicles. SDP includes departmental observations, records, and notices, of unsatisfactory practices related to driving (also differentially weighted).

Analysis. Cronbach's alpha coefficients were derived for each of the HICs described above. Criteria were factor analyzed using an orthogonal varimax rotation. Participants' three individual subscale and ten HIC scores then were correlated with each of the seven criteria and three factors using Pearson's product-moment correlation. Finally, a multiple regression was performed on all the criterion variables and their resultant factors, in which the ten independent variables (HICs) were entered into a regression equation simultaneously.



## Results

The predictor data from 50 of the bus drivers who completed the PEPI were discarded because their Validity subscale scores were not equal to or greater than eight; according to the users' guide provided by the publisher, scores less than eight indicate poor consistency of responding and, therefore, uninterpretable results (Hogan & Hogan, 1989b). In subsequent analyses, some data (N=20) were discarded because participant's identification numbers from the PEPI could not be matched with those from their personnel files. The Ns listed in the following tables also vary according to the number of missing cases associated with each criterion.

Table 2 presents the means, standard deviations, and Cronbach's alpha coefficients of each of the PEPI's subscales and HICs, as well as their intercorrelations. The numbers of items associated with each of the PEPI's subscales and HICs is noted in brackets next to their title.

Table 2. Intercorrelations among Predictors.

														Mean	SD
	1	2	3	4	5	6	7	8	9	10	11	12	13		
1 SO-PA [4]	(.38)	.32@	.14@	.16@	.37@	.22@	.15@	.24@	.32@	.38@	.79@	.38@	.41@	.81	.22
2 SO-UV [5]		(.41)	.11#	.20@	.21@	.20@	.03	.14@	.20@	.11#	.78@	.31@	.18@	.68	.22
3 SO-SE [10]			(.33)	.14@	.17@	-.08*	.04	.01	-.00	.03	.42@	.11#	.04	.89	.11
4 RE-AT [6]				(.63)	.16@	.18@	.10#	.00	.27@	.09*	.24@	.62@	.20@	.83	.22
5 RE-NH [4]					(.51)	.11#	.15@	.08*	.22@	.17@	.37@	.69@	.26@	.68	.29
6 RE-AA [6]						(.57)	.16@	.09*	.35@	.25@	.20@	.65@	.33@	.63	.26
7 ST-NS [2]							(.39)	.04	.19@	.24@	.10*	.16@	.68@	.78	.33
8 ST-CA [2]								(.33)	.15@	.20@	.23@	.10*	.49@	.90	.22
9 ST-NG [5]									(.63)	.37@	.28@	.42@	.70@	.64	.29
10 ST-ND [6]										(.61)	.27@	.26@	.59@	.94	.14
11 SE [19]											(.52)	.42@	.34@	.79	.13
12 RE [16]												(.62)	.40@	.71	.17
13 ST [15]													(.69)	.82	.15

Note: Alpha coefficients noted in parentheses. N ranges from 579 - 634.

Numbers of items associated with each subscale and HIC noted in brackets.

The range of scores possible for any subscale or HIC is 0 - 1.

\*  $p < .05$  #  $p < .01$  @  $p < .001$

Table 3 displays the results of the factor analysis of the criteria. Three factors emerged, accounting for 54.9% of the total variance explained. The first is characterized as "Irresponsibility" because it is heavily loaded by the variables AL, NPR and SDP. Factors 2 and 3 are more difficult to name, since apparently both positive and negative variables load on these in the same direction. For example, PE and HDP contribute approximately equally to Factor 2, and both are positively associated with it. Factor 3 may be more easily characterized, since it might be argued that SRTI, as the self-report of a negative incident, is some index of integrity or honesty and might logically be expected to associate with PPR.

Table 3. Factor matrix.

Criterion	Factor 1	Factor 2	Factor 3
AL	.48	-.44	-.27
PE	-.28	.70	-.13
SRTI	.18	.03	.47
PPR	-.02	-.03	.84
NPR	.85	.12	-.04
HDP	.35	.60	-.04
SDP	.74	.12	.11
Variance explained:	24.8%	15.3%	14.8%
N = 472			

Table 4 presents the means, standard deviations and intercorrelations of the criterion variables.

Table 4. Intercorrelations among Criteria.

	PE	SRTI	PPR	NPR	HDP	SDP	F1	F2	F3	Mean	SD
AL	-.13#	.04	-.04	.30@	.02	.07	.34@	.58@	-.24@	4.99	6.16
PE		-.03	-.02	-.07	.02	-.11#	-.04	.74@	.14@	.51	.38
SRTI			.02	.06	.04	.05	.15@	-.02	.48@	.73	1.25
PPR				-.01	-.02	.01	-.08*	.00	.84@	.18	.36
NPR					-.21#	.52@	.85@	-.16@	.02	.36	.54
HDP						.09*	.52#	.46@	-.02	.28	.82
SDP							.73@	-.11#	.16	.08	.19
F1								.00	.00	.00	1.00
F2									.00	.00	1.00
F3										.00	1.00

Note: \*  $p < .05$ , #  $p < .01$ , @  $p < .001$ . Ns range from 470 to 472.

Table 5 shows the correlations among the predictors and criteria. As can be seen, several predictors and criterion measures were significantly correlated; the strongest relationship observed was a negative one between ST-No Guilt and Hard Driving Performance ( $r = -.15$ ,  $p < .001$ ). The relationships between predictors and criteria are further explored in Table 6.

Table 6 contains the separate regression equations predicting the seven criterion variables and three factor scores from the ten HICs. ST-No Somatic Complaint significantly predicted five criterion variables: Absenteeism (Beta =  $-.13$ ,  $p < .01$ ), Self-Reported Traffic Incidents (Beta =  $.12$ ,  $p < .05$ ), Negative Performance Recognition (Beta =  $-.10$ ,  $p < .05$ ), Soft Driving Performance (Beta =  $-.16$ ,  $p < .01$ ), and Factor 1 (Irresponsibility, Beta =  $-.13$ ,  $p < .01$ ). ST-No Guilt predicted Hard Driving Performance (Beta =  $-.20$ ,  $p < .01$ ), Irresponsibility (Beta =  $-.13$ ,  $p < .01$ ) and Factor 2 (Beta =  $-.12$ ,  $p < .05$ ). RE-Attachment predicted Positive Performance Recognition (Beta =  $.13$ ,  $p < .01$ ), Negative Performance Recognition (Beta =  $.11$ ,  $p < .05$ ), and Factor 3 (Beta =  $.13$ ,  $p < .05$ ). Absenteeism was



## References

- Ash, P. . Baehr, M.E., Joy, D.S., & Orban, J.A. (1988) Employment testing for the selection and evaluation of bus drivers. Applied Psychology: An International Review, 37, 351-363.
- Ghiselli, E.E. (1973) The validity of aptitude tests in personnel selection. Journal of Applied Psychology, 26, 461-477.
- Hogan, J., Hogan, R., & Busch, C.M. (1984) How to measure service orientation. Journal of Applied Psychology, 69, 167-173.
- Hogan, J. & Hogan, R. (1986) Hogan Personnel Selection Series Manual. Minneapolis, MN: National Computer Systems, Inc.
- Hogan, J. & Hogan, R. (1989a) How to measure employee reliability. Journal of Applied Psychology, 74, 273-279.
- Hogan, J. & Hogan, R. (1989b) Hogan Personnel Selection Series User's Guide. Minneapolis, MN: National Computer Systems, Inc.
- Hogan, R. (1990) Personal communication.
- Muchinsky, P.M. (1986) Personnel selection methods. In: International Reviews of Industrial and Organizational Psychology 1986, Cooper, C.L. & Robertson, I.T. (Eds.), New York: John Wiley & Sons, Ltd.
- Schmitt, N., Gooding, R.Z., Noe, R.D., & Kirsch, M. (1984) Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Table 5. Correlations among Predictors and Criteria.

Scale/ HIC	Criteria									
	AL	PE	SRTI	PPR	NPR	HDP	SDP	F1	F2	F3
SO-PA	.04	.03	.07	.02	.03	.06	.07	.06	.06	.06
SO-UV	-.01	-.07	.12#	.02	-.00	-.06	-.00	-.02	-.07	.08*
SO-SE	-.01	.09*	-.02	.04	-.05	-.06	-.05	-.07	.04	.01
RE-AT	.10*	.01	-.01	.02	-.07	-.08*	-.07	-.08*	-.07	-.02
RE-NH	-.12#	-.01	.00	.01	-.05	.02	.02	-.02	.06	.04
RE-AA	-.02	-.09*	.04	.08*	.10*	-.01	.06	.06	-.06	.11#
ST-NS	-.11#	.04	.09*	.05	-.06	.00	-.12#	.09*	.09*	.09*
ST-CA	.03	-.02	.03	.06	.06	.04	-.01	.04	-.01	.06
ST-NG	.01	-.02	.08*	.03	-.05	-.15@	-.01	-.11#	-.09*	.06
ST-ND	-.01	.05	.01	.05	.03	.02	.05	.04	.04	.05
SO	-.02	-.00	.10*	.04	.03	-.01	.03	.02	-.00	.09*
RE	-.04	-.05	.01	.07	-.04	-.02	-.01	-.04	-.02	.07
ST	-.06	.02	.12#	.07	-.04	-.05	-.09*	-.09*	.02	.12#

Note: \*  $p < .05$ , #  $p < .01$ , @  $p < .001$ .

Table 6. Multiple Regression Results (Expressed as Beta) of the PEPI HICs on Criterion Composites and Factor Scores

Predictors	Criteria									
	AL	PE	SRTI	PPR	NPR	HDP	SDP	F1	F2	F3
SO-PA	-.02	.06	.05	.00	.05	.10	.08	.10	.10	.02
SO-UV	-.02	-.07	.11*	.03	.00	-.06	.00	-.02	-.06	.09
SO-SE	.00	.08	-.03	.02	-.04	-.04	-.02	-.04	.03	.00
RE-AT	.14#	.03	-.05	.00	-.09	-.04	-.07	-.08	-.06	-.07
RE-NH	-.13#	-.06	-.08	.00	.00	.05	.04		.04	.01
RE-AA	-.05	-.11	-.02	.13#	.11*	.05	.08	.08	-.04	.13*
ST-NS	-.13#	.02	.12*	.02	-.10*	.00	-.16#	-.13#	.08	.08
ST-CA	.04	.00	.03	.07	.04	.03	-.04	.02	.00	.05
ST-NG	.05	.00	.09	-.02	.07	-.20@	.09	-.13@	-.12*	.00
ST-ND	.00	.06	-.05	.00	.02	.02	.07	.05	.05	-.03
R	.23	.17	.20	.15	.19	.22	.21	.22	.19	.20
R Square	.05	.03	.04	.02	.04	.05	.04	.05	.04	.04
Adjusted R Square	.03	.01	.02	.00	.01	.03	.02	.03	.02	.02
F	2.35#	1.21	1.77	.95	1.53	2.05*	1.99*	2.20*	1.99	1.73

Note: N = 425. \*  $p < .05$ , #  $p < .01$ , @  $p < .001$ .

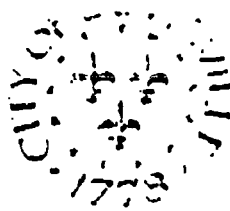
also predicted by RE-Avoids Trouble ( $\text{Beta}=.14$ ,  $p<.01$ ) and RE-No Hostility ( $\text{Beta}=-.13$ ,  $p<.01$ ). The last statistically significant result is that of SO-Unlikely Virtues in the prediction of Self-Reported Traffic Incidents ( $\text{Beta}=.11$ ,  $p<.05$ ), which may contradict the interpretation of SRTI noted for Table 3.

With all the HICs included simultaneously as predictors in the regression equations, four of the ten yield significant Rs. These are in the prediction of Absenteeism ( $R=.23$ ,  $p<.01$ ), Hard Driving Performance ( $R=.22$ ,  $p<.05$ ), Soft Driving Performance ( $R=.21$ ,  $p<.05$ ), and Irresponsibility ( $R=.22$ ,  $p<.05$ ).

### Discussion

The current study failed to replicate a number of results previously reported for the PEPI. For example, Hogan & Hogan (1989a) report that in a study employing 56 combination truck drivers, Reliability was found to correlate with the number of commendations received ( $r=.51$ ,  $p<.01$ ) and with the number of discharges from work ( $r=-.28$ ,  $p<.05$ ). In a second study of 110 line haul drivers (cited in Hogan & Hogan, 1989a) Reliability was also significantly correlated with the number of commendations received ( $r=.15$ ,  $p<.05$ ). The comparable variables in the present study, PPR and NPR, were not significantly correlated with Reliability. In fact, no criteria were significantly correlated with the Reliability scale, although some significant results were obtained using the HICs. Even these, however, cannot be consistently interpreted, as RE-AA is significantly and positively correlated with both PPR and NPR, for example. Similarly, the user's manual published for the PEPI (Hogan & Hogan, 1986), cites a positive correlation between number of commendations received and the Stress Tolerance subscale in the sample of combination drivers referred to above ( $r=.42$ ); the only significant positive correlation obtained for Stress Tolerance in the present study is with SRTI. Service Orientation also did not correlate with any variable but SRTI, but results reported in the user's manual (Hogan & Hogan, 1986b) cite a positive relationship with commendations in the same sample of combination drivers ( $r=.24$ ). Together, the results of the present and previous studies suggest that local validations for this instrument be conducted before its implementation in an employee selection strategy.

In sum, the results of this validation study provide only weak support for the usefulness of the PEPI as a screening device for the selection of the district's school bus drivers. The validities observed in this study are consistent with those reported elsewhere for personality tests and work performance criteria (Ghiselli, 1973; Schmitt, et al., 1984); apparently, a non-pathological basis for scale development did not improve the predictive value of personality assessment in this case. Though some of the HICs are logically and statistically correlated with criterion variables, the lack of strong concordance indicates that the PEPI is of little practical utility in this setting.



# City of Louisville

## CIVIL SERVICE BOARD

609 W. Jefferson Street • Louisville, KY 40202-2728  
(502)625-3565

JERRY E. ABRAMSON  
MAYOR

JERRY W. LEE  
DIRECTOR

### JOB SIMULATION TRAINING & FEEDBACK SESSIONS: THE GOOD, THE BAD, & THE UGLY

By Jeff Prewitt, Chief Examiner  
Louisville Civil Service Board

A paper presented at the International Personnel Management  
Association Assessment Council Conference  
in San Diego, California  
June 1990

**JOB SIMULATION TRAINING & FEEDBACK SESSIONS:  
THE GOOD, THE BAD, & THE UGLY  
By Jeff Prewitt**

This paper presents the pros and cons of applicant training and feedback sessions and describes the current job simulation training and feedback sessions provided for the Divisions of Police and Fire in Louisville, Kentucky.

**THE GOOD**

Properly conducted applicant training and feedback sessions can provide at least four benefits to your organization. The first two benefits discussed below are gained directly by the testing organization and the next two benefits directly affect the candidates. First, these sessions encourage employee acceptance of the testing process. The testing agency can describe and "sell" the testing process and dispel myths and rumors by providing accurate information about the process. Explaining how someone was evaluated and the way they were evaluated may enable you to avoid expensive litigation. This approach can be particularly helpful when you explain the use of the job analysis and SMEs, and that evaluations are based strictly on observable behaviors. Finally, seeking feedback from candidates during the feedback sessions and applying their suggestions will encourage employee acceptance of the testing process and can lead to the second benefit-improved testing procedures.

The feedback sessions can improve testing procedures by providing an opportunity to reevaluate procedures every time they are conducted and by motivating the assessor/developer to adequately justify all ratings. During feedback, the job simulation administrators can seek input on improving the process from the candidates who have just participated. Since one of the assessors will be giving feedback, there is more motivation to accurately document and justify ratings of behaviors during the evaluation process. Using the rating forms for feedback purposes also helps the assessor/developer identify areas for improvement on the rating forms.

The training and feedback sessions help candidates prepare for the testing process and also serve as an employee development tool. With proper training, everyone can know what is expected of them before they attend the job simulation. This knowledge helps build self confidence because the candidates feel better prepared for the testing process. The training and feedback can also be used as an affirmative action tool as it helps motivated minorities to become better prepared.

Not only are candidates better prepared for the job simulation, but the feedback and training can be used as an employee training tool because principles and information from training and feedback can be applied on the job. The feedback sessions can be used as an employee development tool by helping candidates identify strengths, weaknesses and areas in which to seek further education, experience, and/or training. Departmental deficiencies and training needs can also be identified when many candidates exhibit the same weaknesses or deficiencies. This is accomplished through careful and close interaction between SMEs and exercise developers producing not only a good indication of carefully researched desirable job behaviors but also a practical exercise that demonstrates preferred job behaviors to the candidate.

#### THE BAD

Reasons given for not conducting training and feedback sessions vary from concerns with reducing validity and defensibility of the testing process to the argument that a testing division should test and leave the training to the trainers. There is some concern that training and feedback may reduce validity. If candidates know the content and criteria of the testing process, they may adopt a testing strategy which is different from their normal management strategy. This could result in improving test performance without improving job performance. The people who take advantage of the training and feedback sessions may have an advantage over those who do not. You may also invite challenges to scoring procedures or other weaknesses in your process. The last argument is that the role of management training belongs to the training bureaus in the Police and Fire Divisions, not Civil Service.

#### THE UGLY

The ugly side of job simulation training is where an outside consultant or someone within the Division of Police or Fire takes the most motivated candidates, or the ones with enough money, and trains them for success on the job simulation. This is one more reason to conduct your own training program. You should control the training process by insuring the training is available to everyone, and controlling the type of information disseminated.



## **CURRENT TRAINING PROCESS**

Our current training process is a 35 minute videotaped presentation that is made available to everyone regardless of assignment or shift. The Division of Police provides the tape to interested candidates. The Division of Fire plays the tape over their local access television station. The topics covered on the training tape are: history of job simulations; job analysis process; assessing procedures; types of exercises; assessor training and composition; advantages of videotaping exercises; feedback process; and pointers on how to prepare and present yourself. The pointers are the main reason that the candidates watch the tape, therefore, the pointers are at the end of the tape.

## **CURRENT FEEDBACK PROCESS**

The feedback process is divided into two parts. The first part is an overview and request for suggestions for improvement. The second part is specific feedback on each individual exercise.

The overview is conducted by the Chief Examiner. The general feedback given is an explanation of the numbers and the rating process. The candidate is then shown how he/she did relative to everyone else. The Chief Examiner uses this session to relax the candidates and put them into the right frame of mind. Since candidates are often suspicious and defensive and want to try to convince you why they should get higher scores, it is necessary to explain that ratings will not be changed and that they need to have an open mind and listen to their feedback and the reasons they scored as they did. It is explained that this is their time to learn how to do better the next time they face a similar situation whether, it is during a job simulation or on the job. They are also informed that the feedback will be tailored to their needs. In other words, if they wish to review tapes of their performance and get specific feedback, or if they just want to find out what they did poorly, we will accommodate them. This session is concluded with a request for feedback from the candidate. As stated earlier, valuable information can be gleaned at this point. The candidate is also given an opportunity to blow off steam by complaining to the Chief Examiner rather than attacking staff members during the next phase of the feedback process.

The specific feedback is performed by the analyst who coordinated the development of that particular exercise. That analyst was the project coordinator who worked with SMEs in developing the exercises and rating scales. That analyst understands the exercise intimately and was involved in the assessing as the assessor team leader. The use of videotaped performance for feedback gives the candidate an exact record of their specific behaviors. There are typically three exercises on a job simulation so the candidate gets specific feedback from three different analysts.

Job simulations produce a wealth of behaviorally anchored job-related information which needs to be used to its full potential. Candidate training and feedback are two ways to make job simulations more valuable to participants and to the organization. This paper is designed to assist someone considering candidate training and feedback and to stimulate discussion of the pros and cons of job simulation feedback and training so that other agencies can make an intelligent and informed decision on whether or not to implement training and feedback with their job simulations.

*A Shopper's Guide to Personnel Assessment Professionals*

Jeffrey P. Feuquay, Ph.D.  
Assistant Administrator  
Oklahoma Office of Personnel Management

Kay S. Bull, Ph.D.  
Professor of Applied Behavioral Studies  
Oklahoma State University

A paper presented at the  
International Personnel Management Association Annual Conference  
San Diego, June, 1990

*Abstract*

IPMAAC recently sponsored a job analysis of the area of personnel assessment. Using graduate catalogs of 24 Ph.D. granting institutions, with clarification from those institutions where necessary, likely recruitment sources for applicants with the knowledge and skills derived from the job analysis were identified. Sought-after skills were cross-referenced to degree programs in the sampled universities. A number of non-traditional recruitment sources were noted.

Questions or comments may be addressed to Jeffrey P. Feuquay, Assistant Administrator, Oklahoma Office of Personnel Management, Jim Thorpe Building, Oklahoma City, OK 73105

## *A Shopper's Guide to Personnel Assessment Professionals*

Once upon a time, virtually all personnel professionals were generalists; they recruited and screened applicants, wrote and administered tests, set pay and determined what job classifications to allocate specific positions to. With the ever-increasing complexity of the field, generalists are all but extinct. Personnel Assessment, with its focus on highly technical, other-worldly issues, and its extraordinary load of regulations and professional guidelines, may be the most specialized of the specialties. Issues of concern to Personnel Assessment Specialists are little understood and often viewed with considerable skepticism by their co-workers who are dealing with "real world" problems. While their technical orientation may lead to an "ivory tower" reputation, it also has the effect of their being relied upon to perform a wide variety of research, evaluation and analysis functions beyond the traditional scope of assessment. In fact, until recently, there was very limited information beyond the local level concerning what Personnel Assessment Specialists were actually involved in. However, even with that limited information, difficulties in finding candidates had led many Personnel Assessment Managers to the conclusion that the breadth of highly specialized knowledge and skills required to operate in those diverse environments may not be found within the applicants from traditional sources.

In 1983, the International Personnel Management Association Assessment Council (IPMAAC) Board of Directors and then-president Barbara Showers established the IPMAAC Job Analysis Task Force. Chaired by Ronald A. Ash, Ph.D., the Task Force was comprised of a broad cross-section of leaders in the field of Personnel Assessment. They were charged with conducting a comprehensive job analysis for the set of functions falling under the rubric of "personnel assessment specialist". The final report, issued in September, 1988, was based on the analysis of the 435 usable surveys which were returned primarily by public sector personnel assessment specialists from the local, state and federal sectors.

As detailed in that final report, the analysis resulted in the retention of 211 of the original 217 tasks, clustered in 15 task-based job dimensions. The activities described by the 15 dimensions were named:

1. General personnel assessment management and supervisory
2. Technical personnel assessment management and supervisory
3. Non personnel assessment management and supervisory
4. Information exchange and communication
5. Training and education
6. Job analysis, description, and classification
7. Selection procedure development
8. Recruitment and preliminary screening
9. Applicant evaluation and screening
10. Basic test/assessment procedure administration
11. Assessment center development and management
12. Selection procedure validation research
13. General personnel research
14. General data analysis
15. Equal employment opportunity, affirmative action, and related

Table 1 of the report, reprinted here as Appendix 1, provided brief descriptions of the 15 job dimensions and indicated the number of tasks in each dimension.

The purpose of this informal survey was to identify the variety of degree areas from which candidates might be drawn for the role of Personnel Assessment Specialist. While a number of traditional sources came to the fore, the focus of this preliminary effort was the identification of non-traditional sources of applicants, especially of applicants who might have a combination of measurement skills and skills in a complementary area that would have a synergistic effect in an assessment team.

## *METHOD*

Graduate catalogs were reviewed for 24 Ph.D. granting institutions, four from each of six regions of the United States: Northwest, Southwest, North Central, South Central, Northeast and Southeast. Those universities are listed in Appendix 2. The review sought graduate-level course work consistent with the IPMAAC Job Analysis and supportive of the broader role being played by personnel assessment specialists. The few questions concerning specific course content and focus were resolved by direct contact with faculty and department heads at the various institutions. From the review, a list of doctoral degree programs and the courses available through them was developed.

## *RESULTS*

Possibly the most serious problem in analyzing the data obtained was one that personnel assessment specialists face daily if they are involved in evaluating education and experience. That is, the organization of educational institutions, and the program and course contents are highly individualized. This problem in analysis is the very problem that leads to difficulties in finding recruitment sources. Courses were grouped by reported content and name into 20 knowledge/skill areas. Basic areas are necessarily broad and may overlap or encompass the content of the more specific advanced areas. The identified areas are listed below.

### *Knowledge/Skill Areas*

1. Industrial/Organizational/Personnel Psychology; Human Factors; Labor Market & Staffing Analysis
2. Organizational/Systems Behavior, Development and Theory; Program Planning, and Policy & Decision Analysis; Management/Supervision; Industrial Relations
3. Personnel; Human Resources Management, Planning & Administration
4. Basic Statistics
5. Advanced Statistics: Factor Analysis, Scaling, IRT, etc.; Expert Systems
6. Research & Program Evaluation
7. Test Construction; Measurement; Survey Instrument Development
8. Recruitment, Selection & Placement
9. Labor Relations; Collective Bargaining
10. Personnel & Public Administration Law; AA-EEO
11. Wage & Salary Administration, Compensation & Benefits
12. Job/Task Analysis
13. Personality & Psychological Appraisal Techniques & Tools, Psychophysics
14. Behavior Analysis, Personality Theory, Performance Appraisal
15. Guidance and Counseling, Career Counseling & Improving Life Skill Competencies
16. Employment Interviewing
17. Human Resource Development/Training; Adult Education & Learning; Instructional Strategies and Methods; Instructional Technology,
18. Analysis/Improvement of Instruction; Needs Assessment, Planning and Design
19. Oral Communication; Public Relations; Business Reporting and Technical Writing
20. Human Relations; Social Psychology; Group Dynamics

An analogous process was used in the grouping of university departments and majors, and resulted in 12 degree areas. The first four degree areas (A through D) were most often seen housed within colleges of education, others were administratively housed within a variety of colleges. Although administrative reporting relationships were generally based on content of the various degree programs, it became obvious from the number of exceptions that pragmatic and territorial concerns often superseded the logical connections. For that reason, names or key words of the departments are preserved to allow recruiters an appreciation of the diversity, and to facilitate the search for like departments in non-surveyed universities.

#### *Department/Degree Areas*

- A. General Teacher Education; Education Development; Education Policy and Administration
- B. Curriculum and Instruction; Instructional Systems; Curriculum, Instruction and Evaluation
- C. Applied Behavioral Studies in Education; Educational Specialist in Evaluation; Educational Psychology; Psychological and Quantitative Foundations of Education; Educational Foundations, Technology, or Research; Counseling Psychology; Adult and Human Resources Education; Education and Human Development Social Contexts
- D. Occupational and Adult Education; Vocational Technical Education; Trade and Industrial Education; Vocational Teacher and Adult Education
- E. Home Economics Education
- F. Management; Business Administration; Public Administration; Planning, Public Policy and Management; Commerce and Industry; Political Science and Public Administration
- G. Psychology; Industrial/Organizational Psychology; Family Practice and Community Health
- H. Statistics
- I. Economics and Labor
- J. Industrial Relations; Human Resources
- K. Speech and Communications; Public Relations
- L. English as a Second Language

The 20 knowledge/skill areas were cross-referenced to the 12 degree areas in which they were offered. Table 1 shows the percent of the surveyed universities offering the listed knowledge/skill areas through each of the 12 degree areas. Using Table 1, a recruiter will note that applicants with skills in area 16, Employment Interviewing, are most likely to have gained those skills in degree areas D or C, both of which are within the Education arena. Analogous conclusions can be reached for other sought-after skills.

#### *DISCUSSION*

This may be considered only the most cursory of reviews of recruitment sources. It does, however, emphasize the diversity of degree programs which appear to provide graduates with the skills needed to build a comprehensive personnel assessment program. It is hoped that others involved in the search for new personnel assessment specialists will add to the list of sources, and will undertake a thorough survey of university programs. That additional information would best benefit the field if it were also used as a tool for informing universities of the skills we seek for them to provide.



Skill Areas	Departments											
	A	B	C	D	E	F	G	H	I	J	K	L
1	0	0	0	0	0	17	42	0	0	4	0	0
2	12	4	17	4	0	67	33	0	4	8	4	0
3	4	0	8	4	0	67	4	0	8	12	4	0
4	21	0	62	8	4	58	79	4	4	12	12	4
5	4	0	29	0	0	4	21	4	0	0	0	0
6	25	8	50	12	4	29	38	4	0	0	0	0
7	21	8	58	0	0	4	75	0	0	4	0	4
8	0	0	4	0	0	12	8	0	0	8	0	0
9	0	0	0	0	0	25	0	0	0	8	0	0
10	0	0	0	0	0	25	0	0	4	4	0	0
11	0	0	0	0	0	4	0	0	4	8	0	0
12	0	0	0	4	0	0	0	0	0	0	0	0
13	12	4	29	0	0	0	67	0	0	0	0	0
14	4	4	12	0	0	0	42	0	0	4	0	0
15	17	0	33	8	4	0	0	0	0	0	0	0
16	0	0	8	12	0	0	0	0	0	0	0	0
17	17	4	17	17	0	12	12	0	0	12	0	0
18	21	8	42	12	0	0	0	0	0	0	0	0
19	8	0	0	4	0	4	0	0	0	4	8	4
20	0	0	12	0	0	0	4	0	0	0	0	0

Approximate Percent of Surveyed Universities with Departments  
Offering Course Work Applicable to Each Skill Area

TABLE 1

## **APPENDIX 1**

### **Task-Based Job Dimensions Derived from Cluster Analysis of the IPMAAC Personnel Assessment Specialist Task Inventory Data**

(Table 1 of the IPMAAC Job Analysis Report)

**1. General Personnel Assessment Management and Supervisory Activities**

Managing a work unit: Setting unit goals and objectives, assigning responsibilities, monitoring progress, evaluating performance of staff, providing feedback and counseling, serving as liaison with other organizational work units, etc. (23 tasks)

**2. Technical Personnel Assessment Management and Supervisory Activities**

Supervising the development and use of personnel assessment tools, and monitoring their effectiveness from a technical/managerial standpoint; supervising personnel assessment research; evaluating compliance of personnel assessment program with laws and regulations; making recommendations as required. (13 tasks)

**3. Non-personnel Assessment Management and Supervisory Activities**

Supervising programs or projects in areas other than personnel assessment: Planning, budgeting and evaluating programs in general personnel or other areas; supervising areas such as payroll, staffing, classification, etc. (8 tasks)

**4. Information Exchange and Communication Activities**

Engaging in general information exchange, whether orally or in writing: Responding to letters, questions, complaints, phone calls, attending informational meetings, writing routine correspondence, making requests for information, etc. (11 tasks)

**5. Training and Education Activities**

Developing and implementing training programs: Assessing training needs and designing programs in response to them; conducting training programs in personnel/human resource management or other areas. (34 tasks)

**6. Job Analysis, Description and Classification Activities**

Collecting and analyzing job information through a wide variety of job analysis techniques (interviews, panels, brainstorming sessions, questionnaires, etc.); using job information to write job descriptions and make classification decisions. (17 tasks)

**7. Selection Procedure Development Activities**

Developing job-related testing devices (written, oral, interview, evaluations of training/experience, performance tests, etc.); performing content validation research; administering tests and analyzing their results. (21 tasks)

**8. Recruitment and Preliminary Screening Activities**

Planning, developing and implementing recruiting programs: Defining labor market, selecting advertising media, preparing and/or reviewing recruiting materials, writing ads, communicating with job applicants directly, etc. (10 tasks)

**9. Applicant Evaluation and Screening Activities**

Reviewing and screening basic applicant data (application, resume, interview form, background data) and using these data to make basic decisions such as applicant's eligibility, qualifications or rating. (7 tasks)

**10. Basic Test/Assessment Procedure Administration Activities**

Coordinating the administration and scoring of tests. Making arrangements for test scheduling and administration, and scoring/tabulating and/or reporting results. (10 tasks)

**11. Assessment Center Development and Management Activities**

Planning and organizing assessment center activities: Developing assessment center exercises or procedures; selecting, instructing, training and/or briefing test administrators, raters, participants and/or candidates. (6 tasks)

**12. Selection Procedure Validation Research Activities**

Conducting empirical validation research activities: Reviewing research literature and job analysis data; selecting predictor and criterion measures and sample to be studied; conducting statistical analysis to determine validity; writing research reports to document validity, utility, etc. (22 tasks)

**13. General Personnel Research Activities**

Performing activities related to personnel research, other than the research itself: Identifying research topics; writing research proposals; discussing projects and/or findings with colleagues, clients or managers; implementing findings. (4 tasks)

**14. General Data Analysis Activities**

Performing statistical analyses on research data: Designing forms to collect, code or tabulate data; applying computer programs or other means for analyzing the data; and interpreting the results. (8 tasks)

**15. Equal Employment Opportunity, Affirmative Action and Related Activities**

Reviewing assessment procedures for compliance with the laws, regulations and principles of affirmative action and EEO: Reviewing recruitment and testing practices for EEO/AA compliance; investigating or responding to complaints; developing or monitoring EEO/AA plans; counseling supervisors and employees on EEO/AA matters. (17 tasks)

## **APPENDIX 2**

### **Ph.D. Granting Institutions Reviewed (N=24)**

#### **NORTHWEST**

University of Idaho  
University of Oregon  
Washington State University  
University of Wyoming

#### **SOUTHWEST**

Arizona State University  
University of California - Los Angeles  
Stanford University  
University of California - Santa Barbara

#### **NORTH CENTRAL**

University of Minnesota  
University of Nebraska-Lincoln  
University of Illinois-Chicago  
University of Indiana

#### **SOUTH CENTRAL**

Oklahoma State University  
Louisiana State University  
University of Missouri - Columbia  
Texas A and M

#### **NORTHEAST**

University of Rochester  
University of Connecticut  
University of Maine  
City University - New York

#### **SOUTHEAST**

Duke University  
Florida State University  
University of Alabama  
North Carolina State University

13

PAPER

Innovations in Peace Officer Selection:  
The California Highway Patrol Experience

Bob Giannoni  
Tim Gaffney

Selection Research Unit  
California Highway Patrol

June 1990

## PAPER

### Innovations in Peace Officer Selection: The California Highway Patrol Experience

#### ABSTRACT

This paper presents the California Highway Patrol's (CHP) approach and successes over the past three years with developing and implementing a new and more discriminating selection process for State Traffic Officers (STO) with higher job-related screening standards. This new selection process has resulted in a significantly increased Academy pass rate for Hispanic and black State Traffic Officer (STO) Cadets which has allowed the CHP to approach an ethnically balanced STO workforce. For the nine years prior to implementing the new selection process, the screening standards for STO had been continually reduced in an attempt to increase ethnic representation. While the ethnic composition of Academy classes did increase under this past approach, the Academy attrition rate for those targeted ethnic groups of Cadets also steadily increased which negated achieving an ethnically balanced workforce. With the new higher screening standards this trend has been successfully reversed.



## PAPER

### Innovations in Peace Officer Selection: The California Highway Patrol Experience

This paper presents the California Highway Patrol's (CHP) approach and successes over the past three years with a new and more discriminating reading and writing abilities screening process for the entry-level State Traffic Officer (STO) position, STO Cadet. It was possible to implement this new selection process in 1987 because the CHP received fully decentralized "testing" authority for the STO classification. These new higher screening standards have resulted in a significantly increased Academy pass rate for ethnic minority Cadets which has allowed the CHP to more quickly approach an ethnically balanced STO workforce. Before describing the new selection written examination, a brief historical perspective focusing on a major problem not uncommon to peace officer selection, and the one problem most associated with the CHP's prior process, will be presented.

#### Historical Perspective

For the nine years prior to 1987, the CHP had experienced a continual reduction in entry-level screening standards, primarily in the existing written examination, for the STO Cadet position. This reduction was directed by an oversight agency in an attempt to increase the STO classification's ethnic minority representation. While lowering standards did increase the ethnic minority representation of STO Cadets in Academy classes, the Academy attrition rate for those targeted ethnic groups also steadily increased. This increase occurred at a disproportionate rate to non-minority STO Cadets, which negated the CHP's goal of achieving an ethnically balanced workforce.

Starting in 1978, and continuing through 1986, the selection utility of the entry-level STO Cadet written examination was progressively reduced through the following steps. First, what started as an 120 item written examination was ultimately pared down to a 60 item examination without any analysis to support the validity of this action. Further, over this same time period, the examination's more difficult and discriminating items were removed. Finally, the process for setting the written examination's pass point devolved to the point where job-relatedness was not considered. Instead, the sole emphasis was placed on minimizing the examination's adverse effect on ethnic minority applicants when setting the pass point.

By 1986 the STO Cadet written examination's lack of test utility had resulted in a selection process that did not discriminate between those applicants who did, or did not, possess adequate reading and/or writing skills. These skills were necessary for STO Cadets to successfully complete the academic courses at the CHP's 20 week Academy. Once again, the written examination had been continually diluted in an attempt to increase the STO classification's ethnic minority representation. The following example describes the Cadet academic attrition problems experienced at the CHP's Academy due to the reduced entry-level selection standards associated with the low utility written examination.

The last Academy class of 1986 (there are five classes per year) experienced a 50 percent Cadet overall attrition rate (28 of 56). Specifically, the Cadet attrition rates by ethnicity for this class were: Hispanic - 70 percent (7 of 10); Black - 80 percent (4 of 5); Asian - 50 percent (1 of 2); and, White - 41 percent (16 of 39). Clearly, even though years of lowering selection standards did increase the total number of ethnic minority STO Cadet applicants available to be hired into the Academy, this approach also screened-in applicants without the necessary skills to successfully complete the Academy. As a result, Academy attrition rates for targeted ethnic minority groups steadily increased at a disproportionate rate. Regretfully, this was counterproductive to the CHP's goal of achieving an ethnically balanced workforce.

In conclusion, lowering selection standards, specifically on a written examination that is utilized to screen for reading and writing ability, has proven to be extremely unproductive in increasing the STO classification's ethnic minority representation at the CHP. It seems only reasonable that this conclusion can be generalized to many peace officer selection processes where a training academy is one success criterion. It is also important to note how unfair it is to peace officer applicants, regardless of ethnic group, to be administered a written examination that lacks the utility to screen in only those individuals with a reasonable chance of success at the academy. To do so is unfair to all applicants who quit their prior jobs with the intention of starting a new career at an academy yet stand no chance of success because of their inadequate ability.

#### New Selection Process

In 1987, when the CHP received fully decentralized "testing" authority for the STO classification, a new discriminating, high utility, and job-related selection process was implemented for

screening entry-level STO Cadet applicants. This new process relied on the significant strengths of the "Entry-Level Law Enforcement Test Battery" and the new "Essay Test" developed by Richard Honey at the Commission on Peace Officer Standards and Training (POST). The "Test Battery", which is utilized by numerous agencies in California, contains a total of 115 items that measure: Writing Ability - 45 items covering clarity, vocabulary, and spelling; and, Reading Ability - 30 multiple choice items based on reading paragraphs, and 40 CLOZE items based on restoring deleted words. The "Essay Test", which is utilized by only one other agency, is a written exercise that is two or three paragraphs in length. This is a holistically scored exercise where applicants describe, and then explain the significance of, a specified life event.

The CHP had been working with POST since 1978 in an effort to develop predictive validity statistics for the "Entry-Level Law Enforcement Test Battery" as justification to one day use this selection examination for STO Cadet applicants. The nine years of research completed prior to 1987 had clearly shown that the POST "Test Battery" would be valid, of high utility, and statistically significant in its ability to predict academically successful and unsuccessful STO Cadets at the Academy.

After the POST "Test Battery" and "Essay Test" were implemented in 1987 to screen STO Cadet applicants, the CHP has seen a significant decrease in Academy attrition. Specifically, the Academy's overall average attrition rate of 35 percent for the years prior to 1987, has been reduced to an average attrition rate of 12 percent. Of equal importance is the fact that there are no longer any significant differences between the various ethnic groups' Academy attrition rates. Moreover, it should be noted that no other major changes were made to the STO Cadet selection process during this time period. Therefore, these significant reductions in attrition can be directly related to the POST Tests' high utility. In addition, this reduction in Academy attrition has saved the CHP over a million dollars each year by minimizing the monetary loss spent during the selection process on unqualified STO Cadets who ultimately fail the Academy.

### Discussion

Based on the CHP's experience, lowering standards on written examinations designed to screen for reading and writing ability is unproductive when the sole purpose is to increase ethnic minority representation in the workforce. On the other hand, driven by today's political climate which dictates an ethnically balanced peace officer workforce, many agencies are reducing

selection standards in an attempt to increase their ethnic minority applicant pool with the specific goal of also increasing their ethnic minority workforce representation. The CHP's experience indicates that when a police officer selection process includes a job-related training academy as the final success criterion, then lowering selection standards is counterproductive. The preferred solution is to utilize a high utility screening process which reduces Academy attrition across all ethnic groups and moves the agency towards a more balanced workforce.

6/26/90 IPMAAC Symposia

Introduction - by Charles F. Sproule

The purpose of this session is to provide a brief summary and some illustrative examples from a report prepared at the request of the IPMA Assessment Council (IPMAAC) President for the National Commission on Testing and Public Policy (NCOT&PP).

The NCOT&PP is a Ford Foundation funded "blue-ribbon" body of leaders in education, training, human resources development, public and private sector employment, military personnel policy, government and law. The Commission conducted a study of the role of testing in the allocation of educational, training, and employment opportunities. The Commission obtained information for its report by inviting papers, and sponsoring hearings, seminars and meetings.

The NCOT&PP's "mandate" has been to:

- "investigate trends, practices, and impacts of the use of standardized testing instruments and other forms of assessment in schools, the work place, and the military
- recommend improvements in testing that would promote the identification and nurturing of talent, especially among racial, ethnic, and linguistic minorities."

In December 1988, the NCOT&PP invited IPMAAC to submit two papers. Both are being published as Personnel Assessment Monographs by IPMAAC. One is "Recent Innovations In Public Sector Assessment" which was mailed to IPMAAC members in June of 1990. The other, which was prepared by Dr. Joel Wieser, Dr. Nancy Abrams, and Dr. Sally McAttee (the "3P's of IPMAAC") is "Employment Testing: A Public Sector Viewpoint."

The "Recent Innovations" paper was submitted to NCOT&PP in June 1989. It was prepared based on information gathered through:

- a survey of 47 IPMAAC members and a request to all IPMAAC members in the December 1988 issue of Assessment Council News
- a review of IPMAAC Conference proceedings since 1980
- a literature review of public-sector assessment related literature

The "Recent Innovations . . ." paper was authored by Charles F. Sproule. Many IPMAAC members contributed to the paper. Four of the contributors will make presentations today on material they supplied for the paper. The paper was also reviewed by three IPMAAC officers (Joanne Adams, Dr. Nancy Abrams and Dr. Barbara Showers) and by three NCOT&PP reviewers (Kaye Evleth, Dr. Richard Reilly and Dr. Lila Quero) before it was finalized.



The paper presents "state of the art" information on a variety of public personnel assessment methods and innovations including:

#### Selected Assessment Methods

- Minimum Qualifications
- Ratings of Experience and Training
- Biodata
- Structured Oral Examinations
- Work Simulation, Work Sample, and Performance Tests

#### Selected Federal Assessment Innovations

#### Application of Technology to Assessment

- Use of Video
- Use of Computers

#### Use of Test Scores

#### Legal Provisions to Encourage Innovations and Research and Demonstration Projects

#### Employment Testing of Persons with Disabilities, and Employment Programs for the Disabled

The findings and recommendations contained in the report are those of the author, and may not reflect the views of contributors, IPMA, IPMAAC, the reviewers, or the author's employer.

Following is a brief summary of findings from the "Recent Innovations" paper:

#### Brief Summary of Findings

##### Selected Assessment Methods

A wide variety of assessment procedures are used in the public service. There has been an evolution of methods and leaps in our knowledge about them in the past decade.

##### Minimum Qualifications (MQ)

The content validity model has been applied to MQ development. Alternatives to traditional MQ's have been developed.

##### Ratings or Training & Experience

Research has identified methods with higher validity than traditional methods, and which appear less likely to screen out minorities.

##### Biodata

Research shows high levels of validity for a wide variety of jobs and criteria of success with low or no adverse impact. Biodata is beginning to have increased use in the public sector despite the variety of problems associated with this method. An example is the Federal government's use for entry level professional occupations.

##### Structured Oral Examinations

Application of the content validity model has been standard practice in public sector oral assessment. These procedures have recently been "discovered" by private sector practitioners. Recent research shows high levels of validity and low or no group differences.



## Work Simulation, Work Sample & Performance Tests

Wide range of occupations and requirements being assessed with many methodologies. High applicant acceptance, high job relatedness and validity, and low or no group difference when compared to paper and pencil tests.

## Selected Federal Assessment Innovations

These include: revisions in entry-level testing, logic-based measurement, work force quality assessment. The paper contains information on the role of OPM in assessment innovations, reviews of federal hiring and assessment practices, and the problems and contributions of OPM in assessment.

## Application of Technology to Assessment

Reviews the application of video and computer technology to assessment (e.g. job-previews, job simulations, candidate orientation, training of examiners, item banking, rapid applicant tracking and processing, etc). A number of public sector studies have shown high applicant acceptance and low adverse impact for computer and video based assessment.

## Use of Test Scores

There has been movement away from use of tests on an absolute ranking basis. Guidelines for the use of test scores which generally advocate the use of test scores for personnel selection or a category or banding bases have been proposed.

## Legal Provisions to Encourage Innovations and Research and Demonstration Projects

Laws and regulations have been established in a few jurisdictions to allow waiver of employment and training regulations for the purpose of encouraging research, demonstration, and innovations to improve selection and other personnel procedures. Unfortunately, funding for these efforts has been minimal.

## Employment Testing of Persons with Disabilities, and Employment Programs for the Disabled

A variety of efforts exist in the public sector to modify testing procedures so they fairly assess special candidate groups. These efforts are described. IPMAAC had published Model Guidelines for Accommodated Testing of the Disabled.

Some public sector jurisdictions have developed programs to provide jobs, training, and other support to applicants with special needs.

## SUMMARY OF KEY FINDING

Recent research results demonstrate that we have identified assessment methods which can contribute to efficiency, economy and increased productivity, as well as minimize test score differences between majority and minority groups. However, in general, increasing the fairness and flexibility of assessment methods, any improvement, often results in increased cost. Procedures with high validity and low or no adverse impact are usually more expensive to develop and administer, and more time consuming to carry out than traditional testing methods. Public policy makers need to commit the resources which will allow increased use of procedures with high validity and low or no adverse impact.

Some of the possible reasons why the alternative measurement methods identified may have less adverse impact than traditional multiple-choice tests are:

- they allow for diverse responses
- they are often presented in formats other than the traditional paper and pencil format
- the training provided to raters
- the representation of minorities in the rater group

Two different trends evident in the field are the use of tailored job-specific tests based upon the content validity model, and the use of tailored general ability tests for groupings of jobs and the conduct of criterion-related and validity generalization studies of such ability tests.

Insert here

Presentations by symposia presenters

#### SUMMARY OF RECOMMENDATIONS TO NCOT&PP

1. Encourage the expansion of innovation, research and diversity in employment testing improvement efforts. Promote increased funding for such efforts.  
Recommend:
  - clear consistent policy directions
  - increased intergovernmental cooperation
  - expansion of resources applied to assessment improvement efforts
  - adoption of laws and regulations which allow for innovation
  - adoption of guidelines on test modification and accommodation
  - encourage special employment and training programs
2. Increase use of and research on assessment methods which have high levels of validity and low or no adverse impact.
3. Recommend public policy on the appropriate use of tests and test scores. Use of well developed tests should be encouraged. Rigid reliance on test scores only should be discouraged.

## Summary of

### Report of the National Commission on Testing and Public Policy

(NCOT&PP)

#### From Gatekeeper to Gateway: Transforming Testing in America

Report released 5/23/90

#### WHY TESTING MUST BE TRANSFORMED:

- Most concerned with over-reliance on group - administered, paper and pencil multiple choice tests used to allocate initial opportunities in education, employment and the military.
- Many current practices in educational and employment testing stand in the way of efforts to identify and develop talent, and to improve the functioning of key social institutions.
- A test score alone should not present an absolute bar to opportunity.

#### SUMMARY OF RECOMMENDATIONS OF NCOT&PP

1. Reorient testing to promote the development of all human talent.
2. Redirect testing from over-reliance on multiple choice tests to alternative forms of assessment.
3. Use test scores only when they differentiate on the basis of characteristics relevant to the opportunities being allocated.
4. The more test scores disproportionately deny opportunities to minorities, the greater the need to show that the tests measure characteristics relevant to the opportunities being allocated.
5. Tests scores are imperfect measures and should not be used alone to make important decisions. Past performance and relevant experience must be considered.
6. Strategies are needed to hold institutions accountable.
7. Testing must be subjected to greater public accountability.
8. Expand research and development programs to create assessments that promote the development of the talents of all our people.

## SELECTED INNOVATIONS IN A STATE MERIT SYSTEM

by  
Paul D. Kaiser  
Principal Examiner  
NY State  
Dept. of Civil Service

Last year, at the request of IPMAAC, Charles Sproule prepared a report for submission to the National Commission on Testing and Public Policy entitled, Recent Innovations in Public Sector Assessment. As the title indicates, the report summarized a number of the newer approaches to employee selection which have been incorporated into Federal, State, and Local civil service systems. A reading of the report makes clear that we, as the professionals most responsible for developing the tests which governmental employees take, have indeed been busy.

For a variety of reasons (most notably our propensity toward litigation) New York State has had experience with a number of the innovations mentioned in the report. The impetus behind our adoption of new approaches and testing strategies comes from a variety of sources. Certainly an organizational desire to minimize the adverse impact that examinations can display against minorities has fostered the development of alternative procedures. Also, the very real threat of litigation and the concomitant drain on staff and financial resources encourages a solid effort to develop the best (and most defensible) selection systems possible. Finally, a great deal of innovation surfaces from the wellspring of initiative possessed by individual staff members themselves: Their desire to do a better job and willingness to expend the effort necessary to accomplish the different and difficult.

Although comprehensive, the report does not cover the gamut of new practices being adopted in New York. The information presented below will focus only on some of those innovations highlighted in the report. These fall into three main clusters: 1) new assessment methods and formats; 2) the use of the computer in the testing environment both as a test medium and a tool; and 3) alternative approaches to reporting test scores.

Rather than attempt to discuss all the developmental and psychometric issues which surround these recent changes, the following material:

- \*\* describes (where appropriate) those factors which impelled us to attempt a new approach,
- \*\* provides an example of where the innovation was administered,
- \*\* outlines some of the advantages and disadvantages of the innovation,
- \*\* comments on the innovation.

### NEW ASSESSMENT METHODS AND FORMATS

WRITTEN SIMULATION TESTS - These Latent Image examinations were originally developed for use on Correctional Service positions (Sergeant). These titles historically have a high potential for litigation. As such, the need was to create highly content valid examinations which could withstand judicial scrutiny. Since candidate populations were large (7,000+) oral examinations and other measures of higher order cognitive skills were not feasible.

The principal advantages of simulations are (1) they can tap facets of candidate performance, such as judgment, decision-making, and

applied knowledge, which multiple-choice tests usually tend not to do; (2) they can tap many of the same abilities we purport to tap with oral tests, and do so in a more consistent manner; (3) they can tap these abilities for the entire candidate field, since we don't have to restrict ourselves to testing only those who pass the multiple-choice tests, as economics usually forces us to do with oral tests; and (4) for candidate fields of 200 or more, they are more cost-effective than oral tests.

This methodology has more recently been used for such diverse titles as Psychologists, Attorneys, Treatment Team Leaders, Correction Lieutenants, and Police Chiefs.

**Advantages** - Permits greater coverage critical higher level SKAP's (e.g., problem solving & decision making) periods; High candidate acceptance as a content measurement tool (e.g., looks like a job sample problem); High face validity; Has demonstrated less adverse impact than MC format tests; Can replace/complement oral tests or assessment centers; Strong union acceptance since it is objectively scored; Able to be administered economically to large candidate populations in a single day administration; Makes for great press for top level agency administrators.

**Disadvantages** - Complex test development undertaking involving teams of SME's and test development staff meeting over extended time; Costly to print and score latent ink booklets and answer records; Specialized printing skills and equipment needed to print latent ink booklets (or outside contract); Next to impossible to develop comparable forms for retest or alternate test date administrations; Test complexity from a candidate point of view, e.g., special instructions to follow; Little re-use can be made of test material; Requires extensive staff training and not all examiners can do the task; Candidate answer records must be keypunched.

**THE VIDEO TEST FORMAT** - This format was designed to add a non-written test stimulus component to otherwise highly verbal test batteries. In addition, the Video Test format was designed to evaluate candidates' interpersonal and observational skills, as well as their problem solving abilities, which were difficult or impossible to consider in the standard written multiple choice and/or simulation test formats.

Perhaps the two principal features of the video test format are: (1) It has a high degree of acceptability to the candidates as being representative of real "on-the-job" actions; and (2) It presents to the candidate situations which would require a significant amount of reading if these same situations were to be presented in written form (it minimizes the verbal component.).

**Advantages** - Permits greater coverage of critical higher level skills and abilities such as reasoning problem solving, and decision making; High candidate acceptance as a content valid test - looks like a job sample problem; Seems to have less adverse than standard written MC format tests (Preliminary analysis); Can evaluate candidate observational/action analysis skills;

**Disadvantages** - Statistical and analytic paradigms need to be further developed to strengthen item/test reliability measures; Complex test development undertaking involving teams of SME's and extended time periods; Need for specialized video production crew and equipment; e.g., producer, director, actors lighting and video techs, prop manager, video editor, and related equipment; Next to impossible to develop comparable forms for



retest or alternate test date administrations; Need for VCR and video monitors at test locations - Administration arrangements are difficult.

**OPEN-BOOK TESTS** - This examination format has been used for several promotional titles within the Department of Correctional Services. During the job analysis phase of test development we asked how knowledge of a critical rule, regulation, or procedure was applied on the job. Apparently, there exists a body of knowledge which needs not be "memorized" but rather when a situation relating to that rule or directive occurs incumbents refer to the appropriate source. The open book test attempts to improve the content validity of the test and simulate on-the-job behaviors more exactly.

**Advantages** - More job related evaluation of critical knowledge that require or permit on-the-job reference in the use or application of the knowledge; High degree of candidate acceptance (face validity).

**Disadvantages** - Adds time to test administration since candidates tend to look up every answer. Added costs to develop, print, and distribute the Open Book Reference material; May contribute to making the test easier and therefore decrease the contribution of the knowledge component to candidates' final test scores.

**STUDY GUIDES** - The initial use of study guides for promotion candidates sprang from a consultant's recommendation. Briefly, the hypothesis was that minorities may prepare for tests differently from non-minorities, and this difference is not reflected in job performance. By focusing their study activities minority test performance might improve relative to non-minority test performance. Research is currently being conducted to evaluate this hypothesis.

Study guides also have psychometric appeal - If candidates better understand the test task, error variance (caused by anxiety, confusion, etc.) may be reduced.

**Advantages** - May act to decrease the adverse impact on minority candidates; In a litigious situation may show judge that every effort was made to ensure selection process was fair to all candidates; May act to standardize test taking behavior since have better candidate understanding of test task. May reduce error variance.

**Disadvantages** - Time consuming to develop, actual value difficult to quantify; Requires additional pre-test lead time for test development staff; Printing and distribution costs; Difficult to ensure every candidate gets study guide in timely fashion. Can generate complaints.

### COMPUTER BASED TESTS

There are three basic types of computer-based tests which have been used or are under development in New York State:

- 1) conventional written tests administered by computer;
- 2) computer-based performance tests; and,
- 3) computer-based simulations

**CONVENTIONAL TESTS** - The examination for Motor Vehicle Representative III is administered directly in the offices of the Dept of M.V. using their on-line computer terminal system. Candidates are provided the questions by the computer and record their answers directly via the keyboard. Areas



tested include: Coding, Arithmetic, Reading, Proofreading and Keyboard Accuracy. Test items are selected on a pseudo-random basis - no two tests are alike, and, by using a pre-set passpoint with the scoring routines built into the testing program candidates are scored instantly.

**Advantages** - Simulate actual working conditions (test was originally designed to be given on a 'walk-in' basis during the working day in a regular DMV office); Immediate scoring, Inter-filing on list; Quick retest (one week); Face validity - DMV Reps work on these very terminals in the natural course of their jobs.

**Disadvantages** - Requires an on-line computer system; Non-Civil Service agency must assume responsibility for the testing process.

**COMPUTER-BASED PERFORMANCE TESTS** - Like the old typing tests, these tests are designed to assess candidates' skill in utilizing the equipment that they will be expected to use on the job.

**DEMO TEST** - This is a performance test for Data Entry Machine Operators. Operationally the test

- 1) collects candidate biographic data and stores it in a database;
- 2) contains three data entry task components (alpha-, numeric- and alpha/numeric-), each with its own practice part;
- 3) consists of five minute practice sessions followed immediately by ten minute test sessions;
- 4) is timed by an internal clock;
- 5) is self correcting - if candidate gets off track, the program will attempt to find where in the test the candidate is actually working, and adjusts the scoring key accordingly;
- 6) uses an empirically derived scoring standard.

**Advantages** - Simulates the exact job task Decentralizable to remote locations; Immediate scoring.

**Disadvantages** - Equipment needed (PC based)

**DBCLERK** - This is a performance test for Database Clerks, persons who spend a portion of most working days using a computer database. Typical job tasks include adding, deleting and updating records, and generating reports from the database information. It is appropriate regardless of whether a main-frame or PC-based database is used on the job. Operationally, the test:

- 1) collects necessary biographic data from the candidate, stores it in a database; 2) consists of three parts: first, the candidate updates a database from a batch of paper forms (additions, deletions and updating of records are all required; next, the candidate completes paper forms based on data stored in the database and, finally, the candidate is asked questions on line about the entries made on the paper forms.
- 3) there is an internal clock to do all timing;
- 4) if the candidate gets off track, the program will attempt to find where s/he is actually working in the test, and adjusts the scoring accordingly;
- 5) part one of the test is scored on the basis of database integrity. Part three is based on the accuracy of transcription/data entry.

**Advantages** - Directly maps the job task; Is decentralizable to remote locations; Can be scored immediately.

## Disadvantages - Equipment needed

WORDPROCESSING TESTS - These performance tests are designed for word and information processing specialists. The difficulty with word-processing machines and software is that every one is different, and proficiency in one doesn't necessarily mean proficiency in another. While a good wordprocessing operator can be trained from one hardware/software platform to another fairly rapidly, it isn't practical to retrain people before they take a test. Therefore, what we created is a test that can be put up on any hardware/software platform, which requires every candidate to exercise the full range of wordprocessing skills, and where we rate what the candidates produce rather than the way they produce it. Operationally, these tests:

- 1) run on any computer-based wordprocessing machine, including PCs.
- 2) require the candidate to access and manipulate a set of files and to output the product according to certain standards. During the test candidates are required to merge files, use headers and footers, correct typos, insert, delete & move material, handle graphics and format material according to instructions, etc.
- 3) allow promotion candidates to use any hardware/software platform in general use in the agency; however, for open competitive examinations we restrict the platform to one or two major types
- 4) The output from the system is centrally scored by the Civil Service Dept.
- 5) There is an objective rating scale available for each problem.

Advantages - Directly maps the job task; is decentralizable to remote locations; Can be scored easily; Does not require special test administration personnel.

NOTES: Everything the candidate needs is included in the program. Runs on any MS-DOS machine and is distributed on diskettes. The entire exam, including collection of biographical data, practice sessions, etc. takes less than one hour per candidate

COMPUTER-BASED SIMULATIONS - The natural habitat for simulations is the computer. While paper-and-pencil and latent image ink simulations are doable, at best they are slow, awkward, and do not exploit fully the simulation technique. Computers not only have the ability to do the 'page-turning' and 'section-finding' required by the written simulation method, but have the capability of hooking into and controlling optical disk drives and video recorders. They also have astonishing graphics capabilities. The so-called 'arcade' video games that feature cartoon characters are essentially simulations using images stored on optical disk. The technology to computerize written simulations is already available. What is needed is the money to make enough test stations available for timely and convenient testing of the candidate population.

We are currently in the developmental stages of administering a written simulation via computer.

## THE USE OF TEST SCORES

RANK-ORDER SCORING - This is New York State's traditional approach to the scoring of written examinations. Succinctly, after a passpoint is determined, the raw scores of passing candidates are

linearly converted to a 70 to 100 point scale. Final scores are reported in whole number increments, seniority and veteran's credits are added to final scores. Appointing authorities must make personnel selections from one of the top three eligibles willing to accept the appointment.

### RULE OF THREE

**ZONE SCORING** - Zone scoring is a method by which different raw scores (number of correct answers) are assigned the same final test score. For example, on a 90 question multiple-choice test any candidate who answered 84 or more questions correct might be assigned a zone score of 100; candidates who received scores between 78 and 83 might be assigned a zone score of 95; this continues to the passpoint. Typically, seniority credits are added to raw scores and veteran's credits are added to the zone score. All eligibles within a given zone are equally reachable for appointment.

A number of factors can contribute to the decision to zone score a test. The decision takes into account all of the steps in the selection process which preceded and follow the written test, the content of the written test, and the extent to which the written test maps the total job.

- When the job analysis indicates that, within a certain range, a greater amount of the attributes being tested for does not predict a greater ability to do the job, then it is appropriate to combine candidates within that range into groups and to assign them equal final scores.
- When the test cannot measure completely some of the major attributes involved in performing the job, test scores in precise rank order do not represent the rank candidates would achieve if all these aspects of performance were measured. Zone scoring allows consideration of these untested attributes.
- When there is a requirement to match the needs of specific jobs in the class with particular candidate backgrounds and abilities, then scores in precise rank order cannot adequately predict the ability of candidates to perform any one specific job. Zone scoring allows a better match of eligibles and jobs.
- When a statistical analysis indicates that the precision of the test does not support a strict ranking, then zone scoring is appropriate.

The principle advantage of zone scoring is that it allows the appointing authorities greater flexibility in selecting eligibles for appointment. Properly used it may serve to increase the validity of the selection system. For example, a local police agency may ask eligibles to indicate any qualifications they possess above the minimum required and zone scoring may allow them to select individuals who have completed courses in police science. On the other hand, unions tend to object (i.e. sue) zone scoring feeling it opens the door for politics and cronyism and undermines merit system principles.

Statistical arguments in favor of zone scoring frequently cite the unreliability of tests - reporting scores in rank order over-emphasizes the ability of the test to predict job success. Further, the standard error of measurement from the above example may indicate that the candidate with 84 correct is not significantly different from those with raw scores of 89. Arguments against take the opposite tact - grouping scores decreases the reliability of the test and ignores valuable information about the

candidates. Also, the candidate with a score of 84 is more like the candidate 83 than she/he is like the candidate with a score of 89.

RELIABILITY-BASED BAND-WIDTH SCORING - This technique is, in essence, a variant on the zone scoring approach. However, rather than assign a zone score to a given range of raw scores, a reachability band for appointment is established. Assume that an analysis of the reliability coefficients for an examination show that a 7 point band-width is appropriate and the top obtained score is a 94. Any eligible with a score in the range of 88-94 is equally reachable. After all candidates with a score of 94 decline or are appointed, the appointable zone slides so that any eligible with a score between 87-93 is reachable.

This approach addresses the arguments of statistically alike candidates being equally reachable. Also, because rank order scores are reported, the agency can use the list in rank order if it so desires.

New York has considered to using this approach on selected exams. However, since it clearly violates the "rule of three" provisions of State law, we have not yet used this approach to certify an eligible list.

SCORE ADJUSTMENTS - To date, New York has used the score adjustment approach only to eliminate or minimize adverse impact and in response to the pressure of litigation.

We have used the adjustment approach only twice. The first technique was to simply separate the examination's frequency distribution into two distinct distributions - one minority and one non-minority. Each distribution was normalized according to standard statistical procedure and the results recombined. The result was an eligible list with equal proportions of minority and non-minority candidates at each score point in the distribution. The litigation which ensued from this adjustment became known as the Bushey suit. The case began in 1982 and has not yet been resolved.

The second approach to score adjustment was based upon correlation data which was collected on incumbents. That data showed that the test score differences between protected class and non-protected class candidates exceeded differences between these groups in on-the-job performance. The test scores of minority candidates were adjusted upward to reflect the job performance differences. This adjustment resulted in the Puma litigation which commenced in 1985 and continues.

For further information on any of the above issues, feel free to contact Paul D. Kaiser, Principal Examiner, New York State Department of Civil Service, Building #1, The State Campus, Albany, N.Y. 12239 or call (518) 457-5591.



## INNOVATIONS IN PUBLIC SECTOR ASSESSMENT CENTERS

By Dennis Joiner  
Dennis A. Joiner & Associates

The use of assessment centers in state and local government is innovation. The assessment center method had its early roots in military applications and was subsequently adapted and refined for selection of supervisors and managers by large private sector employers in the 1950s and 1960s. Use by state and local government began primarily in the mid to late 1970s and has spread rapidly, particularly as a testing methodology for law enforcement and fire service promotions.

Due to the many civil service requirements and environmental differences in local government employment settings, the assessment center models developed in the private sector are not easily adapted to the public sector without substantial modifications. The typical/classical assessment center model used in the private sector includes:

- three to five days of training for assessors
- all assessors are "in house" managers
- the assessment center program is an ongoing program using the same exercises (content) for 12 or more years without changes
- assessors, once trained, serve (on call) for 12 months to life with no additional training
- usually only one assessor observes a candidate in an exercise
- "solo" assessors are also responsible for being role players while simultaneously observing the candidate
- usually teams of three assessors observe each set of five-six assesseees, usually with no overlap between assessor teams.

Due primarily to local government requirements to use "outside raters" who have no prior knowledge of the candidates and due to concerns regarding "examination security" none of the above listed private sector assessment center characteristics are acceptable or workable in most public sector settings. A typical public sector assessment center is likely to include the following characteristics:

- one full day of training for assessors after reviewing prereading material and followed immediately (the next day) with one day of candidate assessment
- all assessors are borrowed from other public jurisdictions
- the assessment center exercises are developed for use once
- assessors are trained for a one time specific assessment center process and if they serve again they receive an additional day of training on the specific exercises to be used and with all other assessors who will serve on that assessment center

- two assessors independently observe and record each candidate's performance in each exercise
- for intense or complicated roles (such as problem employees or employees with problems) separately trained, non-assessor, role players are used
- usually all assessors in an assessor team of six-eight assessors evaluate all candidates in the assessment center (often 10-12 candidates per day)
- the final scoring process is facilitated by a non-assessor facilitator
- shorter exercises are used (e.g., one or two hour inbaskets versus three-four hours)
- one day of candidate assessment in three or four exercises (versus two days with five or six exercises)
- public sector assessment centers are much more likely to use more mechanical (versus consensus) scoring models which maintain the integrity of the dimension weights determined by the "internal" subject matter experts
- finally, public sector assessment centers tend to be much more job related and face valid. General, "off-the-shelf" exercises are almost non-existent in state and local government applications.

This last point is the source of the most innovation in public sector assessment. That is, because public sector assessment specialists work directly from the tasks of the job in determining the situations to simulate in the test, many unique and highly job related exercises and procedures have been developed. Examples include simulated fire scene and hazardous material situations, fire prevention inspection exercises, formal and informal training exercises, police tactical problems (such as hostage and riot situations), and criminal investigation scenarios.

Many more examples can be found in Charles F. Sproule's recent monograph, Recent Innovations in the Public Sector.



Symposium: Recent Innovations in Public Sector Assessment  
Presentation: OPM's Major Program Initiatives in  
Personnel Research and Development

Presenter: Jay A. Gandy  
U.S. Office of Personnel Management

- I. The Office of Personnel Research and Development is a major office in the Career Entry and Employee Development Group in OPM. The office headed by OPM Assistant Director, Marilyn K. Gowing, has three divisions:

Assessment Services -- headed by John D. Kraft  
Policy and Analysis -- headed by Sandra S. Payne  
Testing Research and Applications -- headed by  
Magda Colberg.

II. The Personnel Research and Development Mission

- A. To conduct basic, applied, and innovative human resource management research.
- B. To assist the Office of Personnel Management in solving operational problems.
- C. To assist the Office of Personnel Management in implementing and evaluating governmentwide programs in all areas of human resource management.
- D. To provide expert advice and assistance to departments and agencies in managing their own human resource programs.

- III. Currently a staff of 45, including 28 psychologists, supervisory psychologists, and 7 professionals in other disciplines.

IV. 67 Years of Service

- A. 1922: OPRD predecessor organization created as Examining Division of CSC.
- B. 1960 - 1982: Personnel Research and Development Center.

- C. 1978 - 1979: PRDC plays leading role in drafting Uniform Guidelines on Employee Selection Procedures and in development of Validity Generalization.
- D. 1980's: Leadership in validity generalization and logic-based measurement research.
- E. 1989: OPRD's traditional role is expanded to encompass all human resource management topics.

#### V. PRD's Leadership Role

- A. Supporting OPM's Mission
- B. Serving Agency Needs
- C. Solving Operational Problems
- D. Representing Excellence in Research

#### VI. Key Current Activities and Research

##### A. Assessment Services Division

##### 1. Federal occupational and Career Information System (FOCIS).

- To assist potential job applicants and current employees seeking career change in matching their own interests and skills with specific occupations.
- P.C. software allows individuals to identify their skills and interests, and to match them with Federal occupations.
- Available now through NTIS.

##### 2. Managerial/Executive Programs

- Review and update of the competencies required and the training offered.
- Evaluation of the Federal Executive Institute curriculum.
- Study of Presidential Rank Award winners.

### 3. Urban Youth Opportunities Programs

- To prepare urban youth for entry-level positions through innovative and intensive training programs.
- Formal training programs and individual career guidance; Saturday Academy.
- In partnership with academic and business communities.

### 4. Flexiplace Studies

- To set up flexiplace experiment and evaluate effectiveness; to determine best way to implement program.
- Can save office space, reduce pollution, save energy, help with child and elder care, assist the handicapped.
- Assess flexiplace program feasibility; evaluate cost, productivity, and employee satisfaction; develop training modules.
- Taking place now in agencies that volunteer; nationwide.

## B. Policy and Analysis Division

### 1. Item Bias (Differential Item Functioning) Research

- To study the extent and causes of differential item functioning in OPM written ability tests.
- Mantel-Haenzel procedure, which matches test takers on ability before looking at test item pass-rate differences.
- Multi-phased program of research; first studies completed (relatively few items identified as problems); some changes implemented in test assembly process.

## 2. Quality Assessment Programs

### a. Applicant and Incumbent Studies

- To build a data base of quality information for applicants and incumbents.
- Exam applicants and special program new hires in all occupations; incumbents in selected critical occupations.
- Standardized, machine-readable qualification inquiry filled out by applicants and new hires; qualification inquiry and job performance measures of incumbents.
- Begun in 1989; phase in to entire population, 1990-1992. Two incumbent studies each year.

### b. Workforce Quality Information Exchange

- To provide a forum for the exchange of research and program information on Federal workforce quality assessment and development.
- Facilitate the exchange of data on quality issues through publications and other fora.

### c. Private Sector Comparison Studies

- To obtain comparative data for establishing quality benchmarks.
- Professional associations, major private sector employers, BLS, Dept. of Education.
- Pilot study with private sector data collection in 1990.

## 3. Career Opportunities 2000

- To provide a coordinated approach for meeting anticipated skilled labor shortages in coming years.
- New or current employees whose aptitude and interests do not correspond to their current career paths or critical government needs.

- Analysis to match skill needs with employee aptitude and interests; implement with innovative employee development and utilization strategies.

### C. Testing Research and Applications Division

#### 1. Design, Construction, and Maintenance of all Federal Written Tests.

- A comprehensive testing program reflecting job analytic findings and state-of-the-art psychometrics.
- Tests used for applicants for about 50 percent of Federal Occupations.
- Tests based on job analysis and constructed using latest psychometric technologies.

#### 2. Improving Test Technologies

##### a. Test Simplification

- To reduce the number of tests and rid batteries of redundant subtests.
- Key test batteries: PAC, Non-PAC, Scientific Aide and Technician, and Apprentice.
- Occupations grouped for testing based on common duties and ability requirements.

##### b. Logic-Based Studies

- To apply logic-based testing methodologies to Federal test batteries.
- All jobs for which a written ability test is used.
- Studies to assure tests sample all job-relevant reasoning processes and are free from extraneous measures.
- Implemented for seven tests batteries; others to be phased in.

c. Innovations in Handicapped Testing

- To design fair and effective methods for examining handicapped applicants.
- Design of technically valid test modifications, e.g., for blind and deaf applicants.

D. Cross-Division Program -- Development and Validation of ACWA Examination

1. ACWA -- Administrative Careers With America.
2. Measuring the whole person -- Exam includes written test plus IAR.
  - Job-relevant written abilities tests.
  - Individual Achievement Record (bio-data measure).
3. Cut score to be set based on job performance.
4. Criterion-related validation and construct studies underway.

VII. Building for the Year 2000 ... and beyond

- A. The Federal government will face unprecedented human resource challenges.
- B. Personnel research will help management to make informed decisions.
- C. PRD plans to conduct necessary basic, applied, and innovative research to meet the challenges ahead.



# **A Statutory Authorization for Selection Experimentation**

Paper for Delivery at

IPMAAC Symposia: Recent Innovations in Public Sector Assessment

**Julie Vikmanis**

Staffing Division Manager

Minnesota Department of Employee Relations

Tuesday, June 26, 1990

## **A Statutory Authorization for Selection Experimentation**

I'd like to begin with a little background to help you put our statute for experimentation in selection procedures into some perspective.

Minnesota is a state of approximately 4 million people, half of whom reside in the Twin Cities metropolitan area with the remainder spread throughout what we call Greater Minnesota. There are approximately 35,000 state employees in Minnesota, again, roughly divided in half between those who work in the metropolitan area and those who work in greater Minnesota. Of this 35,000, approximately 25,000 are full-time, unlimited, classified employees of the executive branch, for whose position the Department of Employee Relations as the central personnel agency of state government must determine appropriate classification and develop and administer Civil Service selection and referral procedures.

Minnesota just celebrated the 50th anniversary of the its Civil Service, having begun in 1939. As one of the older state Civil Services in the nation, it has reached an evolutionary stage wherein we believe we have achieved a comfortable co-existence between a merit system which governs classification and selection, a labor relations system which covers the terms and conditions of employment in sixteen state-wide bargaining units for all but confidential and managerial employees, and where we also believe we have a fairly comfortable fusion of merit system principles and Affirmative Action to achieve a representative work force.

The Civil Service System itself is, we believe, quite flexible. One portion of the law contains over 15 provisions for qualifying and non-competitive appointments which may be used under varying circumstances. We also have one of the more liberal certification provisions in the nation. From exams which have been opened to the general public, we refer the top twenty scoring candidates plus all tied with the twentieth score. If this group does not include members of protected groups (women, minorities and disabled persons) for which the agency filling the position has disparities, we expand the eligible list to include two candidates of each protected group for which the agency is disparate.

With all this comfortable co-existence and such great flexibility, it sounds as if we must have a system which would be generally acceptable to most managers and employees. Not so!

In 1986, our Department of Administration conducted a study of the hiring process in response to an age-old legislative question - why does it take so long to fill a job? The study showed that it took approximately 8-12 weeks to announce and conduct an exam to establish an eligible list from which appointing authorities could typically expect to interview and hire a candidate in another 7-8 weeks. Too long they said, and we agreed.

To solve the problem, the study team recommended a series of actions which we had no trouble supporting - a redesigned, automated system to support the process, additional staff to develop and conduct exams more frequently, and more authority to the Department of Employee Relations to find new ways to speed up the hiring process.

The study team had been impressed with alternatives to standard procedures which we had tried and found successful, such as use of category scoring rather than absolute score ranking for some experience and training exams, use of pass/fail scoring for examinations requiring licensure, developing streamlined appointment procedures for positions in shortage occupations. The study team suggested that such constructive experimentation be encouraged and even more actively pursued.

Since some of our experimentation had been pushing fairly close to the lines of legality within our own Civil Service Law, we decided to use this study recommendation as the vehicle to obtain official sanction for experimentation. Accordingly, in our legislative package to implement the results of the study, we proposed a statutory provision to give the Commissioner of Employee Relations authority to conduct experimental or research projects designed to improve recruitment, selection, referral or appointment processes for the filling of state unclassified positions. The handout material includes a copy of the statute for your review.

Because state employee unions are an effective lobbying force in Minnesota, they were successful in obtaining a meet and confer provision, giving them opportunity just as they would have access to the legislative process for selection matters to at least review in advance the design of any experimental projects (something they have to this point shown little interest in exercising). We included a reporting process to give our Legislative Commission on Employee Relations a comfort level regarding oversight of our experiments and we limited the number of appointments to no more than five percent of the total number of appointments in the preceding year. With such controls, the provision was found acceptable by most legislators, many of whom frankly either didn't understand it or really didn't care and it passed as itself an experimental provision due to sunset in two years.

What it did legally was to give authority to the Commissioner to suspend law, rule and administrative procedure in order to conduct experiments. What it did in practicality was to empower our staff with a new problem-solving tool, which we used judiciously and effectively both to solve particular individual problems and to experiment with new ways to improve the system overall. It also added immeasurably to the improving perception of state managers that the Department of Employee Relations could be a forward-looking operation, actively working to find new ways to operate a more efficient personnel system on their behalf.

In the two years the experimental law was in effect, we conducted nine experiments and filled over 200 positions from these experiments. Most experiments were judged successful and none met with major problems or negative reactions from any affected parties including the applicant public. Some experiments were sufficiently successful that we have proposed legislation to incorporate the provisions of the experiment into ongoing law. We experienced a mild setback when our legislative package to make experimentation a permanent provision of the statute and to incorporate several of the experiments into ongoing statute was derailed in the final hours of our 1989 legislative session - not because of anything dealing with the experiments but because the bill containing the proposal also included some salary setting provisions which became a political football. The

experimentation language and the specific provisions we hope to incorporate into ongoing statute are included in our 1990 legislative package which shows every sign of being passed without difficulty.

I've included in the handout a listing of the specific experiments we conducted during the two year operation of our statute, detailing the reason we chose to experiment and the results of the experiment for your information. I will be happy to discuss any of the experiments that you have questions about. I also have available a more extensive report we made to our legislature for anyone who wants to ask for one after the session.

I think, however, that the specific experiments are not nearly as important as the basic concept. As the result of a cooperative study process, we were able to show:

- that our agency had been responsible in administering state laws dealing with hiring
- that some of these laws (not our staff) made hiring difficult for managers and indeed difficult for us to assist managers to hire efficiently
- that as a responsible agency trying within the statute to experiment, we ought be given even greater latitude to experiment with new ways to improve hiring efficiency, while holding on as much as possible to important merit considerations - a conflict we could be entrusted to understand and balance

Most importantly, the option to conduct experiments has caused our staff to think more innovatively and imaginatively when confronting selection problems and opportunities. We see the statutory language as a reward and a benefit for having done good work in the past and as a unique opportunity to do even better work in the future. Our customers - state managers - have been satisfied with the results and we have proven ourselves to the legislature.

We think it's been a win-win all the way around. If any of you are interested in proposing similar provisions to your legislatures, we would be more than happy to provide information and endorsements in support of your efforts.

JV:dc/1138WPP

**MINNESOTA STATUTES  
LAWS OF 1987, CHAPTER 186, SECTION 14**

**Sec. 14. [WAIVER OF STATUTES, RULES AND ADMINISTRATIVE  
PROCEDURES FOR EXPERIMENTAL OR RESEARCH PROJECTS.]**

The commissioner of employee relations may conduct experimental or research projects designed to improve recruitment, selection, referral, or appointment processes for the filling of state classified positions.

The commissioner of employee relations shall meet and confer with the exclusive bargaining representatives of state employees concerning the design and implementation of experimental and research projects under this section.

Any provision of Minnesota Statutes, sections 43A.09 to 43A.15, associated personnel rules adopted under section 43A.04, subdivision 3, or administrative procedures established under section 43A.04, subdivision 4, is waived for the purpose of these projects. This waiver is limited to no more than five percent of appointments made under the waived provisions in the preceding fiscal year. The commissioner shall report by March 1, 1988, and January 15, 1989, to the legislative commission on employee relations the results of the experimental or research projects.

## Selection Experiments - Minnesota Department of Employee Relations 1988-1989

<u>Nature of Experiment</u>	<u>Classes to Which Applied</u>	<u>Reasons for Experiment</u>	<u>Results/Comments</u>
Offering agencies two options for filling high volume entry clerical classes - 1) Rank order referral from eligible list established via written clerical skills test or 2) direct placement referral from local Job Service following pass/fail performance testing.	Clerk Typist 1 Clerk Steno 1	Satisfy management interest in returning to Job Service placement offered during earlier clerical shortage which resulted in speedier job filling. Permit management two choices for filling the same jobs to also satisfy managers who claim poor service/relations with their local Job Service.	Extremely popular with managers as greatly reducing time to fill positions in high turnover classes. Slightly decreased affirmative action hires. Legislation proposed to make Job Service referral a continuing option.
Doubling the number of candidates referred from long but "overused" lists (from top 20 per vacancy to top 40).	Highway Maintenance Workers	Meet management need for fresh faces and more names from a list which was over two years old with candidates who had been referred multiple times to vacancies in the same locations but still contained over 1,000 names without engendering the work and animosity of reopening the exam or removing non-selected candidates from the list.	User agency pleased with results. Use of list extended by nearly two more years. No candidate complaints. Legislation proposed to make this an option whenever a list of more than 200 candidates has been in effect more than 1 year and referred/certified more than 10 times.



<u>Nature of Experiment</u>	<u>Classes to Which Applied</u>	<u>Reasons for Experiment</u>	<u>Results/Comments</u>
Bringing the number of candidates normally referred from statewide Promotional (employee only) lists (10) to that normally referred from competitive open (public) lists (20).	Executive 2, Office Services Supervisor 2 & 3, Building Maintenance Foreman, Leadworker, Supervisor and Coordinator	Discourage management practice of requesting open (public) exams to simply increase the number of in-house candidates from which they can select with no intention of considering non-employee candidates. Encourage use of statewide promotional lists over agency specific lists to improve employee mobility opportunities systemwide through offering more choice to managers who opt for consideration of statewide lists.  Preserve smaller number (and higher scores) of referrals from agency promotional lists which are most commonly used for classes at lower levels covered by contract which includes a seniority consideration among certified candidates.	Acceptable to all employee representatives consulted. Endorsed by managers. Insufficient appointments thus far to determine whether or not it actually promotes increased statewide mobility. Legislation proposed to make this a permanent option.

<u>Nature of Experiment</u>	<u>Classes to Which Applied</u>	<u>Reasons for Experiment</u>	<u>Results/Comments</u>
Vacancy-specific certification (referral)	Personnel Representative	<p>For class of multiple specialties (classification, compensation, labor relations, staffing and various combinations), score candidate background (for those meeting minimum qualifications) in each specialty area and refer on match of area factor score to individual position requirements rather than combine factor scores and refer top down on total score.</p> <p>Satisfy management need for specialized skills.</p>	<p>Management pleased with improved "fit" of candidates to vacancies.</p> <p>Expect to extend experiment to larger group of classes and positions in EDP field using skills bank approach, focusing on programming languages, environments, equipment used, etc.</p>

<u>Nature of Experiment</u>	<u>Classes to Which Applied</u>	<u>Reasons for Experiment</u>	<u>Results/Comments</u>
Management class selection process focused on rapid identification of best qualified candidates.	Personnel Dir. 4 Personnel Dir. 3 (Conducted twice)	Produce list of best qualified candidates more quickly than typical of normal process requiring more refined rankings of candidates. Limit consideration and scoring effort to top end of candidate distribution for list unlikely to be used for more than one or two appointments (where ranking of lower scoring/lesser qualified candidates a useless exercise producing only opportunity for score appeals). Eliminate wasting time with candidates at low end of distribution of qualified candidates. Make greater use of global professional judgment in the rating process. Emulate to some degree the kind of group "search process" managers are fond of for unclassified and academic positions.	First effort produced negative candidate reaction from those who met minimum qualifications and didn't get on the list. Mitigated in later experiment by more careful wording in announcement/ads. Results were produced faster. Hiring supervisors pleased to neutral. More experimentation needed.

200

210

Nature of Experiment

Development of promotional list from employee records rather than open announcement.

Classes to Which Applied

Associate Warden  
Assistant  
Institution  
Administrator

Reasons for Experiment

Newly appointed Warden with mandate to open new prison in a matter of months requested help in quickly putting together management team to staff the facility. Could not afford to wait for exam announcement and list establishment. Offered and accepted option to have list established from computer-generated group of staff throughout Corrections Department statewide meeting minimum experience qualifications determined by reviewing employee class history.

Results/Comments

Warden delighted and effusive in his praise to others about Employee Relations' flexibility and helpfulness. No adverse reactions from employees.

Though provisional appointment possible, this approach effectively (and openly) replicated the sorting process that Warden would have gone through to select a provisional appointee and permitted him to better "keep faith" with the candidates.

JV:dc/1138WPP

# Improved Scoring for Personnel Tests<sup>1</sup>

J. Bradford Sympson  
Testing Systems Department  
Navy Personnel R & D Center  
San Diego, CA 92152-6800

The two most important characteristics of personnel tests are reliability and validity. Fairness considerations are also important, but a "fair" test that is unreliable and/or invalid is of little value. There are two primary determinants of the reliability and validity of a test: the content of the test and how responses to the test questions are scored. In personnel selection and classification, the content of the test must be related to knowledges/skills that are predictive of on-the-job performance. In personnel training, the content of the test must represent the knowledge/skill domain that is being taught.

Assuming that the content of a test has been determined, for example, by a job analysis in the context of personnel selection and classification, or by preparing a domain specification in the context of personnel training, then the next critical step is to select a method for scoring responses to the questions in the test. An inadequate scoring method can reduce reliability and validity, while a good scoring method can increase reliability and validity. For many years, and in many contexts, the prevalent method for scoring tests has been the number/proportion-correct score. This scoring method has worked reasonably well, but recent psychometric research has shown that more efficient ways to measure examinee knowledge/skill are available.

## *Item Response Theory*

One approach to improved test scoring is based on *item response theory* (IRT) (Lord, 1980). IRT provides a logical basis for improving conventional testing methods and also provides a foundation for the development of new testing methods (e.g., computerized adaptive testing). The principal characteristic that distinguishes IRT from traditional testing theory is the use of mathematical models that can be used to compute the probability that examinees of a given knowledge/skill level will answer a particular test question in a specified manner. At this time, there are two mathematical models that dominate practical applications of IRT: the 1-parameter logistic model (Rasch, 1961) and the 3-parameter logistic model (Birnbaum, 1968, p. 405). Both of these models are based on dichotomous (right/wrong) scoring of test items.

A recent development in IRT is the introduction of polychotomous (i.e., multcategory) models. Polychotomous models have been proposed by Bock (1972), Samejima (1979), Sympson (1981, 1983, 1986), and Thissen and Steinberg (1984). These IRT models offer improved test scoring because they extract information about an examinee's knowledge/skill level from the examinee's wrong answers to test questions, as well as the right answers. Sympson (1986) presented empirical evidence that application of a polychotomous IRT model in the context of computerized adaptive testing (CAT) can increase the alternate-form reliability of adaptive tests.

## *The Problem of Multidimensionality*

A characteristic shared by all of the IRT models mentioned above is that they express the probability of a particular item response as a function of a single knowledge/skill dimension. That is, they are unidimensional IRT models. However, many testing theorists and practitioners feel that very few, if any, tests are truly unidimensional. Bejar (1983) comments as follows:

<sup>1</sup> Paper presented at the 1990 International Personnel Management Association Assessment Council (IPMAAC) Conference on Personnel Assessment, San Diego, June 27, 1990. Preparation of this manuscript, and portions of the research described here, were funded by the Office of Naval Research, Arlington, VA. The opinions expressed are those of the author, are not official, and do not necessarily reflect the views of the Navy Department.

In practice, dimensionality is situation specific. That is, dimensionality is not a property of the items but rather of the responses to items under a specified set of conditions . . . there is growing evidence that unidimensionality is rare indeed even in tests that are meant to be unidimensional . . . (p. 18)

Traub (1983) makes the following observations:

An important question is whether or not responses to items that measure educational achievement, even if the achievement domain has been relatively narrowly circumscribed, are likely to satisfy the assumption of unidimensionality. A consideration that bears on this question is that differences in instruction apparently can create multidimensionality where before there had been unidimensionality . . . In addition to training or educational effects per se, there are at least two other factors associated with the enterprise of assessing educational achievement that will affect dimensionality. One of these is the extent to which the administration of the items is speeded; the other is the extent to which the examinees vary in their propensities to answer by guessing. (pp. 59-61)

These considerations lead Traub to conclude that "responses to items that measure educational achievement seem very likely to violate the assumption of unidimensionality, in which case none of the popular item response models is appropriate" (p. 59). Similarly, Hulin, Drasgow, and Parsons (1983) state that ". . . virtually all real item pools are multidimensional to some extent" (p. 260). The preceding comments suggest that test dimensionality must be carefully considered in any application of an IRT scoring procedure.

Multidimensional IRT models have been proposed, but none of these models has seen much application. This is because some of the models are implausible, some are limited to rarely-used response formats, and some do not currently have a computer program available for fitting the model. Another problem is the fact that fitting multidimensional models can be very time-consuming and expensive, due to the necessity to fit a hierarchy of models in order to identify the data's true dimensionality.

Another consideration in applied settings is the fact that most test users want a single, overall measure of the examinee's knowledge/skill level, regardless of the dimensionality of the test. They do not want multiple scores for each examinee. While multiple scores can be combined, unless the resulting composite score has higher reliability and/or validity than a simple unidimensional scaling, most practitioners will not consider the additional information provided by a multidimensional analysis to be worth the time and effort.

In practical applications, it is not prudent to assume that current unidimensional IRT model-fitting procedures will prove adequate for data that are "not too multidimensional." While there is evidence that unidimensional IRT models can sometimes be used with multidimensional item-response data (Drasgow & Parsons, 1983; Reckase, 1979), these studies also indicate that for some data sets the knowledge/skill dimension that is measured will not be closely related to either the first principal component of the data or the second-order general factor that underlies a set of correlated primary factors. Thus, with current unidimensional IRT model-fitting procedures, there is a danger that the knowledge/skill dimension that is measured will not correspond to the dimension that is of primary interest to the practitioner.

### *Polyweighting*

One way to circumvent the dimensionality problem is to use a scoring procedure that makes no assumptions about the dimensionality of the tests analyzed, while offering higher reliability and/or validity than number/proportion-correct scoring. One such approach is the polychotomous item-scoring procedure that Simpson (1988) calls *polyweighting*. In polyweighting, an empirically-derived scoring weight is assigned to each possible response to a test question. An examinee's "polyscore" is equal to the mean of the scoring weights of the categories chosen by the examinee. Polyweighting does not require the assumptions of IRT, and can be applied with smaller samples than are commonly required with IRT models. Polyweighting does require that item calibration be carried out with a random sample of examinees from the population of



interest.

Polyweighting gives the examinee more credit for correct answers to difficult questions and less credit for correct answers to easy questions. Conversely, polyweighting penalizes the examinee more heavily for wrong answers to easy questions than for wrong answers to difficult questions. This may be contrasted with number/proportion-correct scoring and with scoring under the 1-parameter and 2-parameter logistic IRT models. The latter scoring methods assign scores to examinees in a manner that renders the scores independent of the difficulty of the questions answered correctly or incorrectly (Birnbaum, 1968, p. 458).

In polyweighting, the scoring weights assigned to item-response categories are referred to as "polyweights." An iterative procedure must be used to derive polyweights for a set of items. The procedure is as follows:

- (1) Each examinee in the calibration sample is assigned a provisional score equal to the examinee's proportion-correct among items the examinee was administered. It is assumed that different examinees may have been administered different item-sets during data collection.
- (2) Since proportion-correct (PC) scores for examinees who are administered different item-sets are not directly comparable (due to variation in difficulties and other characteristics of the items administered), each examinee's PC score is converted to a percentile rank relative to those examinees who were administered the same item-set.
- (3) For each item, the mean percentile rank among examinees who chose each possible response category is determined. This computation includes all examinees who were administered a given item, even if they were administered different item-sets.
- (4) Next, for all items and all response categories, provisional polyweights are computed as follows:
  - (a) For each correct answer, the provisional polyweight is equal to the mean percentile rank among examinees choosing the category, rounded to the nearest integer.
  - (b) For each wrong answer chosen by 100 or more examinees, the provisional polyweight is equal to the mean percentile rank among examinees choosing the category, rounded to the nearest integer.
  - (c) For each wrong answer chosen by fewer than 100 examinees, the provisional polyweight is a rounded linear combination of the mean percentile rank among examinees choosing the category and the mean percentile rank among examinees choosing any wrong answer on the item. For these categories, the polyweight for category  $j$  of item  $i$  is equal to

$$W_{ij} = \bar{R}_{i(w)} + \left[ \frac{N_{ij}}{100} \right]^{1/2} (\bar{R}_{ij} - \bar{R}_{i(w)}) \quad (1)$$

rounded to the nearest integer. In Equation 1,  $\bar{R}_{i(w)}$  is the mean percentile rank among examinees choosing any wrong answer on item  $i$ ,  $\bar{R}_{ij}$  is the mean-percentile rank among examinees choosing category  $j$ , and  $N_{ij}$  is the number of examinees choosing category  $j$ .

- (5) Since examinee percentile ranks range from a minimum possible value of  $100(1/N)$  to a maximum possible value of 100, the provisional polyweights can assume any integer value from 0 to 100. For a given item, if the provisional polyweight for an incorrect response is found to equal or exceed the provisional polyweight for the correct response, the polyweight for the incorrect response is set equal to 1 less than the polyweight for the correct response.
- (6) Given the provisional polyweights for all response categories, provisional examinee polyscores are computed. As stated earlier, an examinee's polyscore is equal to the mean of the polyweights of the categories chosen by the examinee. Since polyscores, like all raw test scores, are not comparable between examinees who have taken

different item-sets, the provisional polyscores are converted to percentile ranks within each group of examinees who have been administered the same set of items.

- (7) Given the new percentile ranks for all examinees, the iterative procedure returns to Step 3, above. Steps 3 through 6 are repeated until a convergence criterion is satisfied. In the computer program POLY (Simpson, 1990a), iterations continue until the mean squared correlation ratio between items and percentile ranks stops increasing.

### *Guttman's Scoring Procedure*

Polyweighting is similar to a scoring procedure developed by Guttman (1941). Over the years, Guttman's scoring procedure has been referred to by various names, including *reciprocal averages scaling* (Baker & Hoyt, 1972), *optimal scaling* (Bock, 1960), and *dual scaling* (Nishisato, 1980). Guttman's scoring procedure has the unique property of maximizing coefficient- $\alpha$ , a measure of internal-consistency reliability, for the set of items calibrated (Lord, 1958).

Guttman's scoring procedure has three drawbacks as an approach to test scoring. First, the scoring weights that are derived for an item depend on the difficulty level of the other items that are calibrated at the same time. If an item is calibrated along with a set of easy items, the obtained scoring weights will be different than if the item were calibrated along with a set of difficult items. Second, in order to maximize coefficient- $\alpha$ , Guttman's method sometimes assigns a higher weight to a wrong answer to a question than to the question's correct answer. Third, implementation of Guttman's procedure requires a complete data matrix in which all examinees respond to the same questions. This is a serious disadvantage if one wants to develop and calibrate an item bank that is too large to allow any one examinee to respond to all questions.

Polyweighting avoids these problems. Polyweighting provides scoring weights for an item that are independent of the difficulty of the other items that are calibrated with it. These weights are bounded so that no wrong answer to a question ever gets a higher weight than the correct answer. Finally, data-sets in which different examinees have been administered different questions can be used for item calibration. While polyweighting does not provide absolute maximization of coefficient  $\alpha$ , research has shown that it results in substantial increases in coefficient- $\alpha$ , alternate-form reliability, and domain-related validity.

### *Effect of Polyweighting on Coefficient- $\alpha$ and Domain Validity*

Simpson and Haladyna (1988) conducted an empirical evaluation of polyweighting in the context of medical certification testing. In that study, data from 1100 resident physicians who had completed a 200-item test in the field of otolaryngology (the diagnosis and treatment of ear, nose, and throat disorders) were obtained. Five-hundred of these physicians were selected at random to make up "Sample A." Five-hundred different physicians were selected at random to make up "Sample B." The computer program POLY was applied to the Sample A data in order to obtain summary statistics and polyweights for all 200 items.

Using the set of 200 items as an item bank, Simpson and Haladyna assembled 20 short (10-, 20-, 30-, 40-item) assessment tests and scored them in Sample B. Twelve assessment tests were assembled by randomly selecting items and eight assessment tests were assembled by selecting "best" items. Both proportion-correct scores and test scores based on the Sample A polyweights were computed in Sample B. Then, internal-consistency reliability coefficients were computed and both types of test score were correlated with Sample B 200-item domain scores.

For all 20 assessment tests, polyweighting resulted in higher cross-validated internal-consistency reliability (coefficient- $\alpha$ ) and domain validity in Sample B. The observed increases in reliability corresponded to a mean increase in test length of 28 percent. Over all 20 tests, the mean increase in domain validity was .075. The minimum increase in domain validity was .052.

Results of the Simpson and Haladyna study suggested that polyweighting should allow reductions in test length, while maintaining test reliability at the level observed under traditional

number/proportion-correct scoring. This hypothesis was tested in the study to be described next.

### *Effect of Polyweighting and IRT Scoring on Alternate-form Reliability*

Sympson and Davison (1989) compared polyweighting and dichotomous IRT scoring to traditional number-correct scoring. This study used data collected from applicants for military enlistment who had taken the *Armed Services Vocational Aptitude Battery* (ASVAB). The objective was to determine the impact of these two new scoring methods on alternate-form reliabilities. Content areas studied were ASVAB Mathematics Knowledge (MK) and ASVAB General Science (GS).

Examinees in one part of the study completed one of five 46-item experimental MK tests, and one of six 25-item operational MK tests. These examinees were assigned to either an MK joint-calibration sample (N=6447) or one of 30 MK holdout samples (total N=6434). Using the joint-calibration sample, all 380 MK items were calibrated simultaneously in a single run of the computer program POLY. These items were also calibrated using an IRT item-analysis program that fits the 3-parameter logistic model (Wingersky, Barton, & Lord, 1982).

In the MK holdout samples, each experimental test and each operational test was split in half in order to create five 23-item experimental alternate-form pairs and six 12-item operational alternate-form pairs. (The first item in each 25-item operational test was omitted.) For each alternate-form pair, at each possible test length, alternate-form reliabilities were computed for number-correct (NC) scores, for polyscores, and for IRT ability estimates based on the dichotomous 3-parameter logistic model.

Experimental and operational MK tests were analyzed separately. For each type of test, at each possible test length, median alternate-form reliability was determined for each scoring method. Proportionate reductions in test length that would be possible under polyweighting or under IRT scoring, without reducing alternate-form reliability below the levels observed under NC scoring, were then determined.

Other examinees in the study completed one of four 57-item experimental GS tests and one of six 25-item operational GS tests. These examinees were assigned to either a GS joint-calibration sample (N=5412) or one of 24 GS holdout samples (total N=5398). Using the joint-calibration sample, all 378 GS items were calibrated simultaneously in a single run of the computer program POLY. These items were also calibrated using the IRT item-analysis program.

In the GS holdout samples, each experimental test and each operational test was split in half in order to create four 28-item experimental alternate-form pairs and six 12-item operational alternate-form pairs. (The first item in each experimental and operational test was omitted.) For each alternate-form pair, at each possible test length, alternate-form reliabilities were computed for NC scores, for polyscores, and for IRT ability estimates based on the dichotomous 3-parameter logistic model.

Experimental and operational GS tests were analyzed separately. For each type of test, at each possible test length, median alternate-form reliability was determined for each scoring method. Proportionate reductions in test length that would be possible under polyweighting or under IRT scoring, without reducing alternate-form reliability below the levels observed under NC scoring, were then determined.

The results of this study are summarized in Table 1. Entries in that table show the mean (over all test lengths from 2 items to the end of the test) reduction in test length that would be possible under polyweighting and under dichotomous IRT scoring. Based on their analyses, Sympson and Davison reached five conclusions: (1) Polyweighting is superior to NC scoring; (2) Dichotomous IRT scoring is usually superior to NC scoring, but not always (see Operational GS); (3) When dichotomous IRT scoring works well, polyweighting and dichotomous IRT scoring allow similar reductions in test length; (4) Polyweighting and IRT scoring both provide greater benefits when item difficulties are more variable (as they were in this study's Experimental tests); (5) Polyweighting and IRT scoring use different (complementary) mechanisms for increasing measurement precision.

Table 1  
Mean Proportionate Reductions in Test Length

Type of Test	Scoring Method	
	Polyweighting	IRT Scoring
Experimental MK	13%	15%
Operational MK	8%	9%
Experimental GS	20%	21%
Operational GS	10%	-1%

Sympson and Davison speculated that the failure of IRT scoring when applied to operational GS tests was due to the multidimensional nature of those tests and their relatively short length. They also speculated that the most effective scoring method to use would be polychotomous IRT scoring. However, it seems that an item calibration procedure such as the one described by Sympson (1986) would have to be used in order to avoid problems with multidimensionality.

### Conclusions

Available evidence indicates that scoring methods superior to number/proportion-correct scoring are now available. Dichotomous IRT scoring and polyweighting provide similar increases in test reliability, but IRT scoring may fail when its assumptions are not satisfied. If one wants to implement either dichotomous or polychotomous IRT scoring, it seems safest to fit the IRT model using a procedure that makes no assumptions about the dimensionality of the items calibrated (e.g., Sympson, 1986, 1990b). If one does not have access to samples of the size required for IRT item calibrations (at least 1500 examinees per item, preferably 2500), then polyweighting can be used. Research to date indicates that polyweighting will allow reductions in test length of at least 10-25%, depending on the content area and the spread of item difficulties in the tests. The Sympson and Haladyna study indicates that polyweighting can safely be used with 500 examinees per item. Further research is needed, but I speculate that we will find polyweighting is superior to number/proportion-correct scoring with samples sizes as small as  $N = 100$ .

### References

- Baker, F. B., & Hoyt, C. J. (1972, April). *The relation of the method of reciprocal averages to Guttman's internal consistency scaling model*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver: Educational Research Institute of British Columbia.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.
- Bock, R. D. (1960). *Methods and applications of optimal scaling* (Research Memorandum 25). Chapel Hill, NC: Psychometric Laboratory, University of North Carolina.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.



- Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-192.
- Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *Prediction of personal adjustment* (Bulletin 48, pp. 321-348). New York: Social Science Research Council.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory*. Homewood, Illinois: Dow Jones-Irwin.
- Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, 23, 291-296.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto, Canada: University of Toronto Press.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 4, 321-334.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.
- Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research Report 79-4). Knoxville: Department of Psychology, University of Tennessee.
- Simpson, J. B. (1981, October). *A nominal model for IRT item calibration*. Talk given at the Office of Naval Research Conference on Model-based Psychological Measurement, Millington, TN.
- Simpson, J. B. (1983, June). *A new item response theory model for calibrating multiple-choice items*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.
- Simpson, J. B. (1986, August). Extracting information from wrong answers in computerized adaptive testing. Paper presented in B. F. Green (Chair), *New Developments in Computerized Adaptive Testing*. Symposium conducted at the annual meeting of the American Psychological Association, Washington, DC.
- Simpson, J. B. (1988, May). *A procedure for linear polychotomous scoring of test items*. Paper presented at the Office of Naval Research Conference on Model-based Psychological Measurement, Iowa City, IA.
- Simpson, J. B. (1990a). *POLY: A computer program for polychotomous item analysis* (Release: 4/27/90). Unpublished manuscript, Navy Personnel R&D Center, Testing Systems Department, San Diego. (Available from the author)
- Simpson, J. B. (1990b, June). *Fitting 5-parameter logistic ICCs*. Paper presented at the Office of Naval Research Conference on Model-based Psychological Measurement, Portland, OR.
- Simpson, J. B., & Davison, M. L. (1989, March). *Reducing test length with polychotomous scoring*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

- Sympson, J. B., & Haladyna, T. M. (1988, April). An evaluation of "polyweighting" in domain-referenced tests. Paper presented in C. E. Davis (Chair), *New Developments in Polychotomous Item Scoring and Modeling*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans.
- Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika*, 49, 501-519.
- Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver: Educational Research Institute of British Columbia.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.



## Biographical Data: The Past Predicts the Future<sup>1</sup>

Mary A. Quenette, David G. Ward, Thomas Trent & Gerald J. Laabs  
Testing Systems Department  
Navy Personnel Research & Development Center<sup>2</sup>

The commonly held view that people tend to behave in a consistent manner over time and situations has found expression in the use of the background data questionnaire for selection and placement. As Owens (1976) has stated, "One of our most basic measurement axioms holds that the best predictor of what a man *will do* in the future is what he *has done* (emphases in original) in the past." Cascio (1978) concurs, "Compelling evidence exists that when appropriate procedures are followed, the accuracy of personal history data as predictors of future work behavior may be superior to any known alternative." Interest in biographical data (biodata) dates back at least as far as the turn of the century, and biodata has many successful applications to its credit. (See Reilly & Chao, 1982, and Mumford & Owens, 1987, for reviews of the literature.) Among its many advantages, Cascio (1976) cites ease of development, low cost, potentially high predictive validity, and an easily accessible data base.

### Problem

Military research and development on biographical data (biodata) has been driven by an attrition problem combined with the pressure to cease using high school diploma status as a screening device.

First-term attrition increased dramatically following the inception of the All-Volunteer Force in 1973. In fiscal year 71, while the draft was still in effect, the 36-month attrition rate was 26%, but by 1974, it had risen to 37% (Nelson, 1986). A 1979 study by the General Accounting Office (GAO), which covered fiscal years 1974 through 1977, estimated the total cost of attrition to be \$5.2 billion dollars, or \$11,700 per attritee. This figure includes the costs (estimated at \$2.7 billion) of lifetime benefits, for which approximately half of the personnel were qualified. (Figures are not adjusted for inflation.)

A breakdown of the attrition rate by education level provides insight into the relationship between the two: There are sizeable differences in the rates as a function of education level, with holders of a regular high school diploma attriting at about half the rate of personnel without a diploma (approximately 25% as opposed to about 50%). Holders of alternate degrees have an attrition rate (about 45%) which is very similar to individuals who have no degree. Complicating the picture still further, there has been a proliferation of types of alternate diplomas in recent years, including attendance-based credentials, certificates of attendance, correspondence course work, home study, occupational school certificates and test-based certificates issued by the states. Holders of some of the alternative degrees appear to have an attrition rate similar to regular high school graduates; others, similar to high school dropouts.

---

<sup>1</sup>Paper presented to the 14th Annual Conference of the International Personnel Management Association Assessment Council at San Diego, CA, June, 1990.

<sup>2</sup>The opinions expressed in this paper are those of the authors, are not official, and do not necessarily represent those of the Navy Department.

Three categories of diploma status (regular high school graduate, alternate degree or credential, and no degree or certificate) are used as a primary enlistment screen along with the Armed Forces Qualification Test (AFQT). Of the two main selection devices, educational credential is the better predictor of successful adaptation to military life ( $r_{pbis} = .16$ ). AFQT is limited in its capacity to predict attrition ( $r_{pbis} = .08$ ) being more useful for predicting trainability and job performance (Sellman, 1989). Therefore, attrition control is exercised by bringing into the Armed Forces an overwhelming majority of recruits who have a regular high school diploma... Biodata holds promise for increasing the predictability of attrition regardless of diploma status.

While diploma status has been useful in controlling attrition, heated controversy surrounds the categorization of some of the alternate credentials on a lower enlistment priority level than high school graduates. In recent years, in fact, there has been considerable pressure to eliminate high school diploma status as a determinant of eligibility for enlistment. The American Council on Education (ACE), for example, in reference to the adaptability screening program (GED Items, 1989), states that "We believe that the screening system should *completely* (emphasis in original) replace the system based on educational credentials and that this should occur this year." This belief is consistent with the view that recruiting policy based on educational credentials is a misuse of state-issued educational evaluations and policy should instead be based on individual measures of persistence. A biodata questionnaire would allow enlistment decisions based on individual applicant characteristics, rather than decisions based on group membership. Recruiters have been severely restricted in terms of the allowable percentages of recruits drawn from the pool of applicants who have an alternative degree or no degree at all. Biodata holds promise for identifying many young people who would serve their country well, but are not now allowed to do so for lack of a high school degree.

The Armed Services Applicant Profile (ASAP) was developed in response to the costly problem of attrition among first-term enlistees in the Armed Forces, and also to reduce reliance on high school diploma status as an enlistment screen.

### Test Development

The original item pool for the ASAP was drawn from The Recruiting Background Questionnaire (RBQ) from the Navy and Marine Corps (Atwater & Abrahams, 1983) and the Military Applicant Profile (MAP) from the Army (Eaton, Weltin, & Wing, 1982). Two forms, 130 items each with 90 items in common, were administered to 120,175 applicants for active duty between December 1, 1984, and February 28, 1985, 55,675 of whom enlisted and were tracked through 36 months of service. The enlistees' responses to the items plus knowledge of their success or failure on the criterion, service completion, provided the basis for the scoring key.

Half of the enlistees responding to each form ( $N = 13,685$  and  $13,172$ ) were randomly selected to serve as key construction groups; their responses to the 90 common items were combined ( $N = 26,857$ ) to obtain the greatest possible stability in the keys. The remaining examinees became the cross validation groups ( $N = 12,760$  and  $12,388$ ).

The scoring keys were developed by the horizontal percent method, commonly used for weighted application blanks (Guion, 1965), in which each option is weighted by the percent of respondents choosing that option who are also successful on the criterion. The percent weights were transformed into a three-point scale (1, 2, or 3) by dividing the frequency distribution of weights into thirds; the highest one-third of the weights, indicating a greater probability of service completion, were assigned a score of 3, the middle group of weights were transformed to a 2, and the lowest one-third, associated with a lesser probability of service completion, were

given a score of 1. The percent cutoff points which divided the weight distribution of the common items were then used to transform the percent weights for each option on the remaining unique items. A respondent's score is the sum of the transformed weights assigned to the options selected by that respondent.

Extensive rational and statistical analyses were conducted to identify items which may have given the appearance of bias, or were, in fact, statistically biased (Wise, Hough, Szenas, Trent, & Keyes, 1989) and those items were discarded. Additional items were eliminated if they were intrusive, focused on conditions beyond control of the applicant, were directed toward determining if the applicant had a high school diploma, appeared to be biased against applicants who were economically disadvantaged, or if the empirical keying resulted in option scores which were irrational given the item content. A total of 31 items from the original pool of 170 were eliminated.

Administration time limitations dictated abbreviated versions. Previous research (Trent, 1987a) and a pilot test (Barnes, Gaskins, Hansen, Laurence, Waters, Quenette, & Trent, 1989) indicated the optimal questionnaire length to be 50 items. Each item was then evaluated in terms of overall validity, subgroup mean scores, subgroup validities, and face validity, with the best items selected for inclusion in the final two 50-item forms of the ASAP. Twenty-one items appear on both forms; the remaining items were assigned to either form such that the forms are balanced with respect to subgroup means, subgroup validities, overall item validity and content areas.

The content of many of the items is similar to the type of information usually asked of civilian job applicants, for example, previous or current employment and academic achievement. Other items ask the applicant about interests, social interactions, or delinquent behaviors. ASAP item content areas are consistent with factor analytic results reported in the biodata literature: Academic achievement, nondelinquency, work orientation, social adaptation, career orientation, work ethic, and athletic involvement are frequently reported.

### Criterion

The criterion, service completion, was based on length of time in service as well as reason for separation. Personnel who completed 36 months of service or obtained an early release were designated successful on the criterion and were assigned a score of 2 (64.8% of total enlistee sample). Those who were separated for performance or behavioral reasons prior to 36 months or completion of their contract, were placed in the low criterion group (unsuccessful) and assigned a score of 1 on the criterion (26.0% of total enlistee sample). Finally, active duty personnel with less than 36 months on duty and persons who separated for reasons largely beyond their control (e.g., medical disability, death, hardship) were excluded from statistical analyses (9.3% of total enlistee sample). The overall success rate, i.e., the base rate, was 71.4% (calculated using only the "successful" and "unsuccessful" groups as described above), while 28.6% attrited.

### Properties of ASAP

Many of the prominent objectives of test construction were achieved in the development of the ASAP. Three major achievements are:

1. A level of overall validity which would make an important contribution to increased retention rates for all enlistees,

2. An increase in predictive precision over and above the prediction of attrition provided by the current screens, AFQT and high school diploma status, and
3. An instrument which does not adversely affect members of racial/gender subgroups applying for enlistment.

### Validity

Validity and cross validity coefficients (the point-biserial correlation between test scores and the dichotomous criterion in the key construction and cross validation groups) are shown in Table 1. The large key construction sample produced highly stable scoring keys, as is demonstrated by the trivial loss of validity when the keys were applied to the cross validation samples. The coefficients are of a magnitude similar to those found in the biodata literature. For example, Reilly et al. (1982) report an average cross validity of .30 for tenure in the military.

Table 1

#### Validity and Cross Validity of the ASAP

Form	Group	
	Key Construction	Cross Validation
A	.27	.26
B	.26	.25

The critical question, of course, is whether or not this instrument can improve on prediction of attrition over and above the current screening procedures. The ASAP is regarded as a noncognitive measure; nevertheless, it correlated moderately with both AFQT and diploma status (point-biserial correlation coefficients ranging from .32 to .38). With incremental validity analyses, however, significant and important increases in predictive validity became apparent when ASAP was entered into the regression equation after diploma status and AFQT; the multiple correlation increased .09 and .07 for the two forms. Clearly, ASAP is telling us something about an individual's propensity to complete military service that is not predicted by the other measures.

What all of this means in "bottom line" terms is that the addition of ASAP to enlistment procedures is expected to result in substantial savings in attrition related costs. For example, a utility analysis using the Taylor-Russell approach (Trent, Quenette, Ward & Laabs, 1990) projected a savings of more than \$100 million annually if 10% of applicants, those at the low end of the ASAP score distribution, were declared ineligible for enlistment.

### Fairness

A great deal of effort was directed toward developing an instrument that would not place any racial or gender subgroup at a disadvantage; i.e., unfairly restrict a subgroup member's opportunity to serve in the Armed Forces. As mentioned above, each item was carefully

scrutinized prior to inclusion in the final forms, with particular attention being paid to subgroup statistics. All racial/gender subgroups have higher mean scores on the ASAP than do white males, and throughout the entire range of possible cutting scores, larger percentages of each subgroup would be accepted for enlistment as compared to white males or to the total applicant sample. (Analyses were conducted for white, black and Hispanic females and white, black and Hispanic males. Other groups were inadequately represented, and, thus, statistical analyses were inappropriate.) Differential validity and/or differential prediction were apparent for some subgroups, yet the evidence clearly shows that adverse impact would not occur. Bias analyses are reported in more detail in Wise et al. (1989) and Trent et al. (1990).

### Reliability

Homogeneity in scale construction is generally not emphasized with empirically-keyed biodata instruments and little information on reliability has been reported. Typical internal consistency estimates, however, hover around .70 (Mumford et al. 1987) and ASAP reliabilities are .76 and .74 for the two forms, using coefficient alpha estimates. It should also be noted that Shaffer, Saunders & Owens (1986) have reported that both objective and subjective biodata are reliable from a long term test-retest perspective.

### Alternate forms

A welcome consequence of the test construction procedures employed was that the forms were equated during development, obviating the need for using statistical equating procedures. In Figure 1, the cumulative score distributions for the two forms are plotted: they virtually coincide. Indeed, a study by Waters (1989) concluded that statistical equating would introduce additional error.

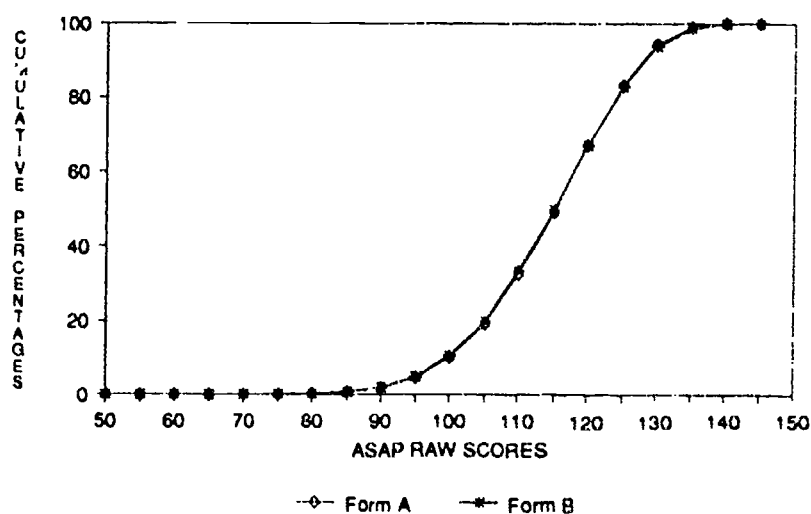


Figure 1. ASAP Score Cumulative Percentages  
for Form A and Form B



## Implementation Issues

All of the testing materials have been produced, including the test booklets, answer sheet, administration manual, scoring instructions, and scoring templates. Unfortunately, in this time of rapidly changing Defense Department budgets, operational implementation has been indefinitely postponed. Several critical issues in test implementation, including coaching/faking, generalizability and the impact of cutting scores, have become particularly compelling in this application of biodata to military selection.

### Coaching/faking

An issue with which users of biodata instruments must deal is the problem of response distortion. The literature on this issue is complex and not easily summarized, but a few general conclusions are warranted. Research results generally support the contention that verifiable items are unlikely to be falsified (Cascio, 1975) and warnings reduce the falsification of items in general (Schrader & Osburn, 1977). While experimental evidence shows that respondents are capable of intentional distortion when motivated (Schrader et al., 1977), Mumford et al. (1987) have observed that there is support for the claims of item accuracy when there is no motive for faking. In a study which has important implications for the ASAP, Trent (1987b) experimentally manipulated ASAP items judged to be susceptible to distortion, yet concluded that the military applicants studied "did not exhibit unrestrained response distortion; distortion was relatively reserved."

The use of an empirically derived scoring key, as is the case with ASAP, is a first step toward reducing intentional distortion, since the "correct" answer is not always readily apparent to the examinee, at least not to the extent that it would be if rational keying were used. Applicants' efforts to represent themselves in a manner which does not reflect their true characteristics, beliefs, accomplishments, or behaviors can degrade validity. In a military selection situation, there is the additional possibility that recruiters, under pressure to meet their recruiting quotas, will coach the applicants.

In anticipation of such a situation, several monitoring plans have been developed, falling into the general categories of deterrence, detection and correction of response distortion (Hanson, Hallam, & Hough, 1989). Deterrence of distortion, discouraging the applicants from distorting their responses in the first place, takes the form of a strong warning contained in the ASAP instructions regarding falsification and the consequences (i.e., possible denial of enlistment). ASAP items vary in the extent to which they may be confirmed: About one-fourth are verifiable and an additional one-fourth are at least partially verifiable. Given the large number of applicants tested each year (over half a million), verification on more than a random basis represents an enormous logistical problem. The value of this type of item in a military selection situation lies in the applicant's perception that his/her answer could (or will) be checked.

Not relying on warnings alone, however, a score monitoring system has been designed to detect inflation of scores above the mean score baseline of the trial administration. Statistical comparisons of operational mean score levels with the baseline will be conducted at regular two week time intervals (Waters & Dempsey, 1989). Divergence beyond chance level will result in closer investigation, with appropriate corrective action taken.

Finally, research has been conducted into developing formulas to correct scores if distortion has been detected (Hanson et al., 1989). This approach has been less fruitful. At this time, it is not clear that this would be a feasible solution although it has been used in other settings and the validity of content scale scores has improved (McKinley, Hathaway, & Meehl, 1948).



### Generalizability

The second concern is generalizability. Wise et al. (1989) discuss the attenuation of validity which may occur with empirical scoring keys, but conclude that the ASAP scoring system should generalize to new samples. Further, a comparison of the ASAP sample with the applicant and enlistee populations from the two most recent fiscal years instills confidence. The differences in demographics are minor and overall the ASAP sample parallels the more recent populations.

### Cutting Scores

A third issue involves the setting of cutting scores on the instrument. Although recent changes in the world are expected to have an impact on recruiting, the recruiting climate has been unfavorable in the recent past. The Services have been experiencing difficulty in obtaining applicants with good qualifications, and are therefore reluctant to turn away applicants they would have accepted based on the current screens. The complexity arises in attempting to integrate the use of the ASAP with the AFQT and educational level. A cutting score or scores must be set to maximize the retention rate under the constraints of the recruiting situation and still satisfy manpower needs. Each branch of the military sets its own standards for enlistment, and a cutting score policy appropriate for one branch may be inappropriate for another. Although the screening out of some candidates would initially require additional time, effort and money to meet quotas, in the long run, successful prediction of attrition will enhance retention rates, lower demands on recruiters, and save a substantial amount of money.

## References

- Atwater, D. C., & Abrahams, N. M. (1983, December). Adaptability screening: Development and initial validation of the Recruiting Background Questionnaire (RBQ) (NPRDC Tech. Rep. 84-11). San Diego: Navy Personnel Research & Development Center.
- Barnes, J. D., Gaskins, R. C., Hansen, L. A., Laurence, J. H., Waters, B. K., Quenette, M. A., & Trent, T. (1989, March). The Adaptability Screen Profile (ASP): Background and pilot test results (IR-PRD-89-06). Alexandria, VA: Human Resources Research Organization.
- Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. Journal of Applied Psychology, 60, 767-769.
- Cascio, W. F. (1976). Turnover, biographical data, and fair employment practice. Journal of Applied Psychology, 61, 576-580.
- Cascio, W. F. (1978). Applied psychology in personnel management. Reston VA: Reston Publishing Company, Inc.
- Eaton, N. K., Weltin, M., & Wing, H. (1982, December). Validity of the Military Applicant Profile (MAP) for predicting early attrition in different educational, age, and racial groups (TR-567). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- General Accounting Office (GAO) (1979). High cost of military attrition can be reduced (FPCD-79-28). Washington, DC: Author.
- Guion, R. M. (1965). Personnel testing. New York: McGraw-Hill.
- Hanson, M. A., Hallam, G. L., & Hough, L. M. (1989, November). Detection of response distortion in the Adaptability Screening Profile (ASP). Paper presented to the 31st Annual Conference of the Military Testing Association, San Antonio, TX.
- McKinley, J. C., Hathaway, S. R., & Meehl, P. E. (1948). The MMPI: VI. The K scale. Journal of Consulting Psychology, 12, 20-31.
- Military adaptability screening. (1989, May/June). GED Items. Washington, DC.
- Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. Applied Psychological Measurement, 11, 1-31.
- Nelson, G. R. (1986). The supply and quality of first-term enlistees under the all-volunteer force. In W. Bowman, R. Little, & G. Sicilia (Eds.), The all-volunteer force after a decade. Washington, DC: Pergamon-Brassey's.
- Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand-McNally.
- Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.
- Schrader, A. D., & Osburn, H. G. (1977). Biodata faking: Effects of induced subtlety and position specificity. Personnel Psychology, 30, 395-404.

- Sellman, W. S. (1989, November). Implementation of biodata into military enlistment screening. Paper presented to the 31st Annual Conference of the Military Testing Association, San Antonio, TX.
- Shaffer, G. S., Saunders, V., & Owens, W. A. (1986). Additional evidence for the accuracy of biographical data: Long-term retest and observer ratings. Personnel Psychology, 39, 791-809.
- Trent, T. (1987a, August). Armed forces adaptability screening: The utility of the biographical inventory. Paper presented to the 95th Annual Convention of the American Psychological Association, New York City.
- Trent, T. (1987b, August). Armed forces adaptability screening: The problem of item response distortion. Paper presented to the 95th Annual Convention of the American Psychological Association, New York City.
- Trent, T., Quenette, M. A., Ward, D. G., & Laabs, G. J. (1990). Armed Services Applicant Profile (ASAP): Development and validation (in review). San Diego, CA: Navy Personnel Research and Development Center.
- Waters, B. K. (1989, April). ASP 01A and 01B equating. Paper presented to the Joint Services Selection & Classification Working Group, Arlington, VA.
- Waters, B. K., & Dempsey, J. R. (1989, November). Development of the Adaptability Screening Profile score monitoring system. Paper presented to the 31st Annual Conference of the Military Testing Association, San Antonio, TX.
- Wise, L. L., Hough, L. M., Szenas, P. L., Trent, T., & Keyes, M. A. (1989, September). Fairness of the Armed Services Applicant Profile (ASAP) Final Report. Washington, DC: American Institutes for Research.

## Computer Based Instruction Technology

*Douglas Weizel, Ph.D. †*  
*Navy Personnel Research & Development Center*  
*San Diego, CA 92152*

Compared to conventional instruction, the greater amount of effort required to create good Computer Based Instruction (CBI) becomes more noticeable when attempted by developers with lower levels of expertise in the 'art' of this technology. Of course a programmer can develop CBI from a technical standpoint, but that person may have no background in the specific content of the instruction to be developed, and neither may have experience in instructional matters. Further, the range of instructional delivery situations varies considerably and the difficulty of presenting instruction in a computerized format has its own unique considerations. Thus, good CBI generally requires either a team or a person with a combination of subject matter, computer and educational skills. To compensate for the diverse set of skills required, there is a need for CBI 'authoring' systems that encapsulate instructional development expertise and easy human-computer interfaces that only leave the content expert out of the equation. To the degree that similarities in instructional delivery techniques can be identified, it is possible to develop specialized interfaces that home in on these situations and make them easier to accomplish. The specialized interfaces described here employed generative CBI techniques for learning situations with similarities from instance to instance that allowed those instances to be generated from previously unassembled components. This work was generally targeted for low cost microcomputers used by people without high levels of programming expertise.

### User Experience and Task Difficulty

To understand the difficulty of producing CBI, at the minimum we need to view the situation as a matrix of user experience level, the difficulty associated with creating different types of instructional delivery, and the difficulty of the instructional content itself. Quantifying these dimensions immediately suggests other dimensions within each, for which there is no immediately agreed upon taxonomy. First, a gross picture of the experience situation is suggested by the progression from users (a) with no computer experience at all, (b) users with some computer experience in operating programs from prompts or manipulating files or knowing about file system organizations, changing directories and word processing or text editing experience, (c) users with some experience in some 'authoring' system for CBI or even some elementary control or programming experience with .BAT files or BASIC or the like. Second, a gross picture of the difficulty of the instruction might contain elements about: whether videocdisc production is associated, whether only a quizzer is desired, whether feedback is desired, whether an accompanying tutorial is desired, whether the instruction is homogeneous enough to permit a database sampling approach, whether simple to complex graphics are involved, where foils for multiple choice items come from, whether alternative answers are accepted if typed in answers are given, ...etc. Lastly, yet a third dimension affecting the ease of making the instruction is the degree of technical familiarity with the content domain to be taught.

### Specializing for Common Situations

While an obvious first step in reducing the task difficulty for instructional 'authors' of varying backgrounds is to create easier interfaces, further steps rely on specializing the application for selected instructional domains. The identification of routine instructional situations which can have specialized interfaces adapted to them requires careful analysis so that sufficient flexibility is available for varying instructional content and user preferences. To the extent that the expertise of more seasoned computer based instruction developers is "canned" in

---

† The opinions expressed are those of the author and do not reflect those of the Department of the Navy.

the programs, the difficulty for less experienced CBI authors will be reduced. That is, by isolating common instructional strategies, we can routinize them so that many details need not be attempted, such as screen design or methods for receiving answers from students. A further advantage is that we can reduce the "rough edges" of the products created by the inexperienced by retaining some control over the allowable interfaces. In constraining the instructional delivery options, the loss of freedom to the developer can be minimized to the extent that provisions are made for the many variations in a specialized instructional domain. Some factors favoring and not favoring routinization of an application are summarized below in Table 1.

Table 1. Routinization Factors

---

Factors favoring routinization (little author input or overhead):
Is standard student interface possible?
Is standard database of questions and answers possible?
Is standard method of giving a tutorial possible?
Is standard method of giving feedback possible?
Is standard method of process control possible?
Factors demanding unique creation or not favoring routinization:
The text of the question & correct answer **
Specifying alternative correct answers
Specifying unique feedback for correct answer
Specifying incorrect answers & their feedback
Specifying a tutorial presentation **
Non-linear process control for advancement & branching
Variable manipulation & storing states
** (Must be input regardless of being routine or unique)

---

The factors favoring routinization mean that there is a lessened amount of authoring overhead. That is, the author does not have to invest as much effort to input information and to configure the presentation to the student. Without a specialized interface, these factors could require considerable effort with a conventional frame based authoring program. The factors not favoring routinization involve greater authoring overhead and demand the creation of unique instructional configurations or process control features. Items marked with asterisks ( \*\* ) are the core instructional content that that an author would have to uniquely input whether or not a specialized routinized interface were available. The other factors here are ones brought into the instructional development process as it matures and demands more uniquely tailored student interactions.

One approach to routinizing instructional presentations detailed here is to provide "templates" for routine or common instructional situations. Templates for question and answer frames are found in some authoring packages to save steps in creating instruction with a number of similar frame sequences. These partially completed templates are duplicated over and over in order to fill in the information unique to each. A generative CBI approach advances beyond this elaborative application by generating new instances for a single template. Thus, just one template is used again and again, with new instances inserted each time by some higher level algorithm. The instructional content must be cast as some form of a structured database so the algorithm can access its components. The algorithm inserts the components into prearranged slots in the template which are 'dynamic placeholders' or variables. Such slots might be prearranged spots on the screen for text, graphics, or answers. The entry of instances into the slots may in turn involve subordinate algorithms specific to the prompting question, the analysis of the answer, and the delivery of various forms of feedback. The technique generally depends



upon some degree of similarity from instance to instance in the features of the interface for the tutorial, question, answer, and feedback.

### Instances of Generative CBI

Three instances of generative CBI techniques are described below. Two of them represent applications in specialized instructional domains. The third is applicable to a variety of situations and incorporates greater control over the interface and presentation algorithm.

(1) **Generating Questions from a Semantic Network:** A semantic network has been used to represent large bodies of facts that are to be memorized by examining the facts and by receiving quizzes on facts or picture recognition. Questions and foils are generated on-the-fly from database assertions as the student programs run, avoiding the need for authors to create large numbers of question screens. The semantic network consists of *assertions* that have three parts: *Subject-Relation-Object*. These assertions may be used either organizationally to subcategorize items in the database with a default *isa* relation, or more frequently to create unique relations that assign object attributes to the subjects of the assertions. Figure 1 shows that having assigned information to the three parts of the assertion then permits the information to be presented to the student in one of three ways: (a) Statements of the assertion as would be shown during database familiarization in preparation for a quiz. (b) A question seeking the *object* as an answer, given the formatted prompt presenting the *subject* and *relation* of the assertion to the student. (c) A question seeking one or more *subjects* as an answer, given the formatted prompt presenting the *object* of the assertion to the student. The formatting of these three presentations is accomplished via the three templates shown in Figure 1, where the format statements contain dynamic placeholders for the *subject* (*%s%*), *relation* (*%r%*) and *object* (*%o%*). At run-time during the execution of the student program, these dynamic placeholders are replaced by the current *subject*, *relation* or *object* as the template is shown to the student.

Figure 2 illustrates a question screen shown to the student, and three potential screens resulting from selecting an answer mode. When there are several answers sought to a single question, the interface indicates the number of answers required. To gain the most number of points, answers may be typed in, with alternative spellings being accounted for in an answer analysis scheme. For fewer points, students may pick answers from a Multiple Choice list, with the incorrect foils being selected from related 'siblings' of the correct answer to make them maximally difficult. These foils are automatically generated according to several rules controlling how many there are and where they come from, and extra effort must be made to create unique foils for a multiple choice item. Students lose points for incorrect answers, or if they simply ask to be told the answers with a 'tell-me' mode. Incorrect answers invoke a scheme that increases the probability that missed questions will be asked again at a later time.

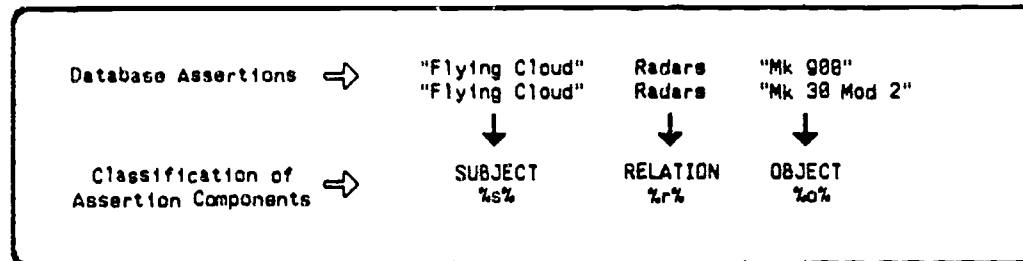
Several programs use this basic theme in different ways. One program lets student browse the database to examine the facts via the statement templates, or to view any associated graphic/video displays of an item. Another program shows associated graphic or video displays and then asks for identification of the image using the answer format described above. Another program combines the features of the above programs in a minimal sort of tutorial for students unfamiliar with the domain. This was accomplished by administering a sequence in which an item is first described by showing its picture and listing all of the factual attributes of the item, and then administering a fact quiz on just those descriptions with the method previously described in Figure 2.

The situations to which this technique lends itself are generally ones where a large categorized set of terse facts must be memorized. Examples of the databases that have been used: ships or aircraft and their attributes (e.g., radars, weapons, equipment, and associated attributes such as speeds and ranges), cranial nerves, and picture quizzes (e.g., electrical symbols or flags). Some databases have yielded as many as 3000 questions for the bi-directional type of questions illustrated above.



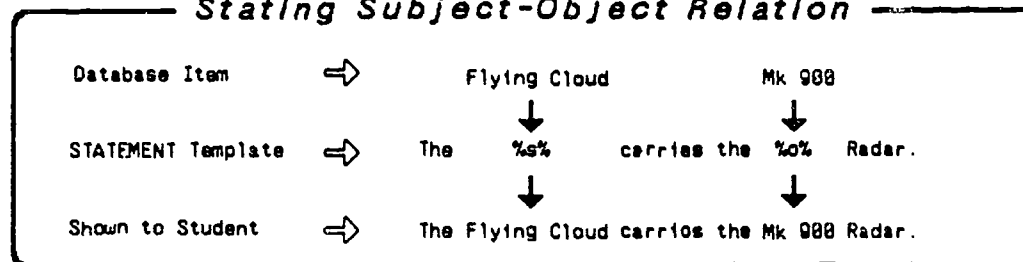
Only the final program described above provided a minimal student tutorial, by using an elementary presentation manager to show all pictures and facts before giving a quiz on the facts shown. Thus, the theme of this package is generally just for memorization training. These programs all include a standardized student interface, database and feedback.

### Database Contents

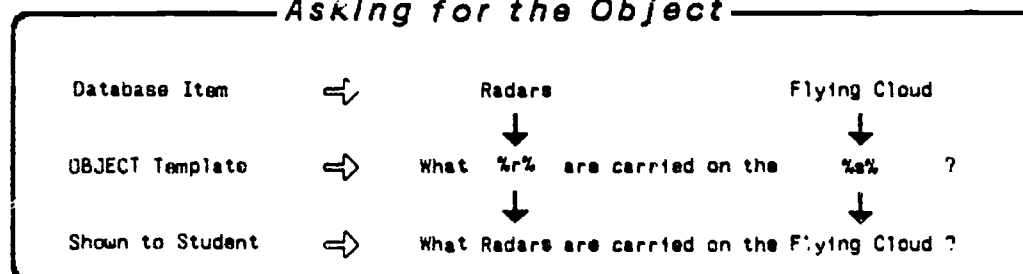


### Presentations to Students Via Three Template Transformations of the Database Assertions

#### Stating Subject-Object Relation



#### Asking for the Object



#### Asking for the Subject

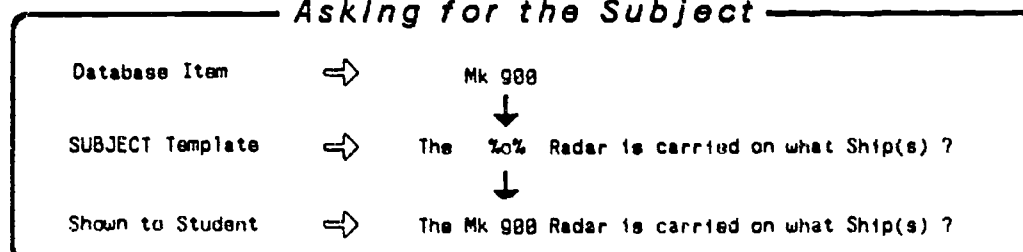


Figure 1. Transforming database assertions into statements to students.

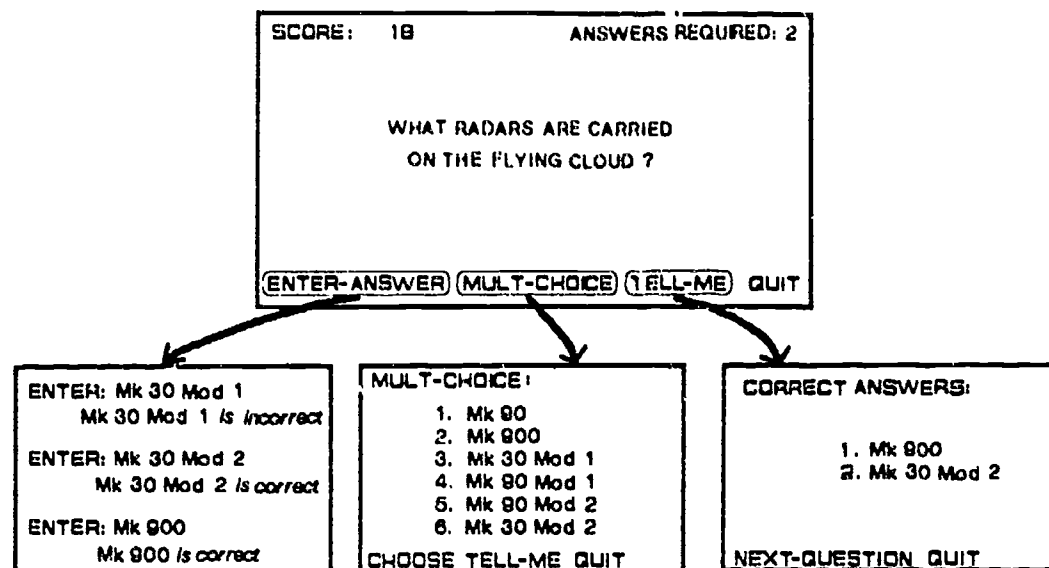


Figure 2. Students enter answers by typing them or by multiple choice, or can simply receive answers with tell-me.

(2) **Technical Vocabulary:** A well circumscribed domain with routine presentation methods makes technical vocabulary training an ideal circumstance for the application of generative CBI. Given a database of technical terms and their definitions, the use of templates and standardized presentation methods creates a circumstance where student learning activities can be rapidly created. A program implementing these techniques was developed to automatically generate student activities that include: spelling, multiple choice, true/false, matching, definition building. Figure 3 illustrates definition building, where a student serially selects the next phrase in a definition. The definition is initially broken into five phrases by the program (or author) and the student is shown the first phrase as a prompt and asked to select the next phrase from a multiple choice list of phrases that are taken from the definitions of other words in the exercise. Foils automatically created in this manner are used in the other student activities, such as the conventional multiple choice test where incorrect answers are drawn from definitions of other words designated by the author to be tested in the exercise. While incorrect foils do not have to be created uniquely for each question, the requirement for creating them in special circumstances is not accommodated. Thus, this is an instance where an automated template technique reduces the difficulty of creating the presentations, but at the same time removes some flexibility and freedom from the developer.

The automatically generated activities will all present learning/testing sessions when a minimum of a word and its full definition have been entered into a database with a menu driven authoring program. Enhancements to these presentations are possible by entering optional additional word data. For example, while the full definition is used if all else fails, a word may have several definitions, as well as special definitions better adapted to just the definition building and matching activities. The presentation or feedback for a given word may also be tailored by designating associated words (tagged as synonym, antonym, related word), by designating words confused with other exercise words (with a sentence to distinguish the two), and by a general feedback to be presented when a wrong alternative has been chosen during multiple choice and true/false activities.

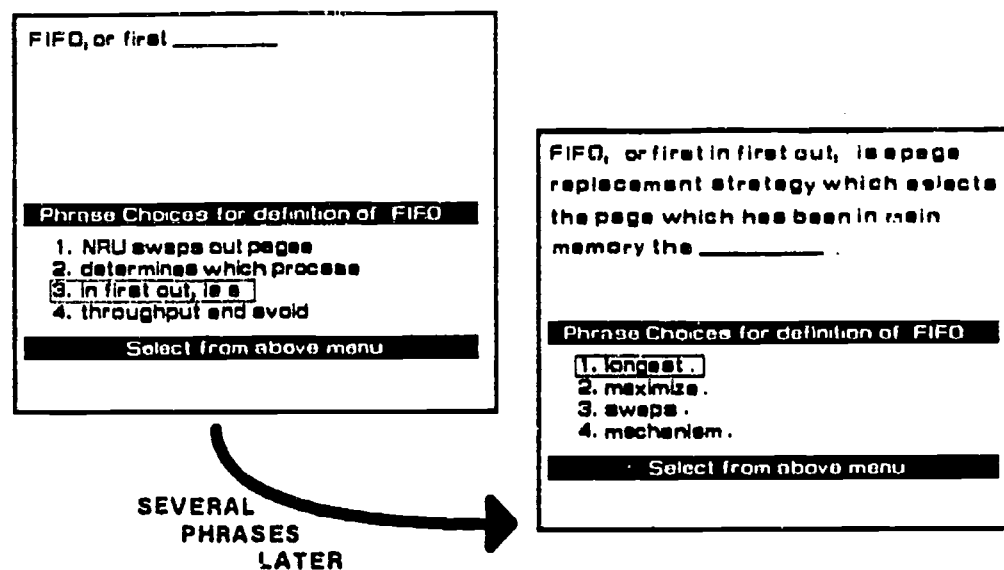


Figure 3. Building a vocabulary definition phrase by phrase.

A second level of specification is provided in the program so that the presentation can also include an initial tutorial and activities that cannot be generated automatically. These manually created activities include an introductory sequence that can be reviewed before beginning the learning/testing activities, and two free form activities known as cloze paragraph & labeling. The cloze activity is a paragraph in which designated words are replaced with blanks for the student to complete. The labeling activity contains blanks for the student to complete which are positioned around a graphic or textual display. Taking all activities together, there are a mixture of both routine and unique instructional presentation techniques. A typical lesson might consist of an introductory sequence of instruction with content that partially lends itself to quizzing the definitions of words, but also contains relations among those words. The relations among words could be concepts that are not easily quizzed in the automatically generated formats, so the final portion of the lesson might consist of a labeling or cloze paragraph activity with missing words to test pieces of the concept.

Figure 4 illustrates the relations among the word definition database, exercise specification and the activity presentation algorithm. The *activities* are the previously discussed learning/testing methods and *exercises* are one or more named collections of those activities. The first step in the authoring process is to enter words, their definitions, and other word definition data. The second step is to name an exercise and select words from a menu that are to be included in the current exercise. At this point the author may create a unique tutorial, cloze or labeling activity for this exercise. The third step occurs when the student program is executed. The author arranges to have the program executed with command line option letters that correspond to the subset of learning/testing activities desired. Multiple exercises may be executed back to back to create longer or unique combinations of presentations. The learning/testing activities selected require some judgement on the part of the instructional developer since not all activities are appropriate to all types of content.

The program has been used for a variety of instructional contents, such as electrical/electronic terminology, navigation & maneuvering terminology, reviewing basic study skills, and familiarization with components of valves. Such bodies of instruction generally include a substantial number of terms and the automated use of templates in this specialized application has been successful in reducing the authoring overhead and has been used by developers with lower levels of experience. In some circumstances the program is used for remedial training, and in others it is used for initial familiarization with a domain prior to

beginning more in depth study that integrates the vocabulary in other concepts. When the focus of the instruction goes beyond the limitations of the program, then more flexible authoring programs are required, such as described in the next section.

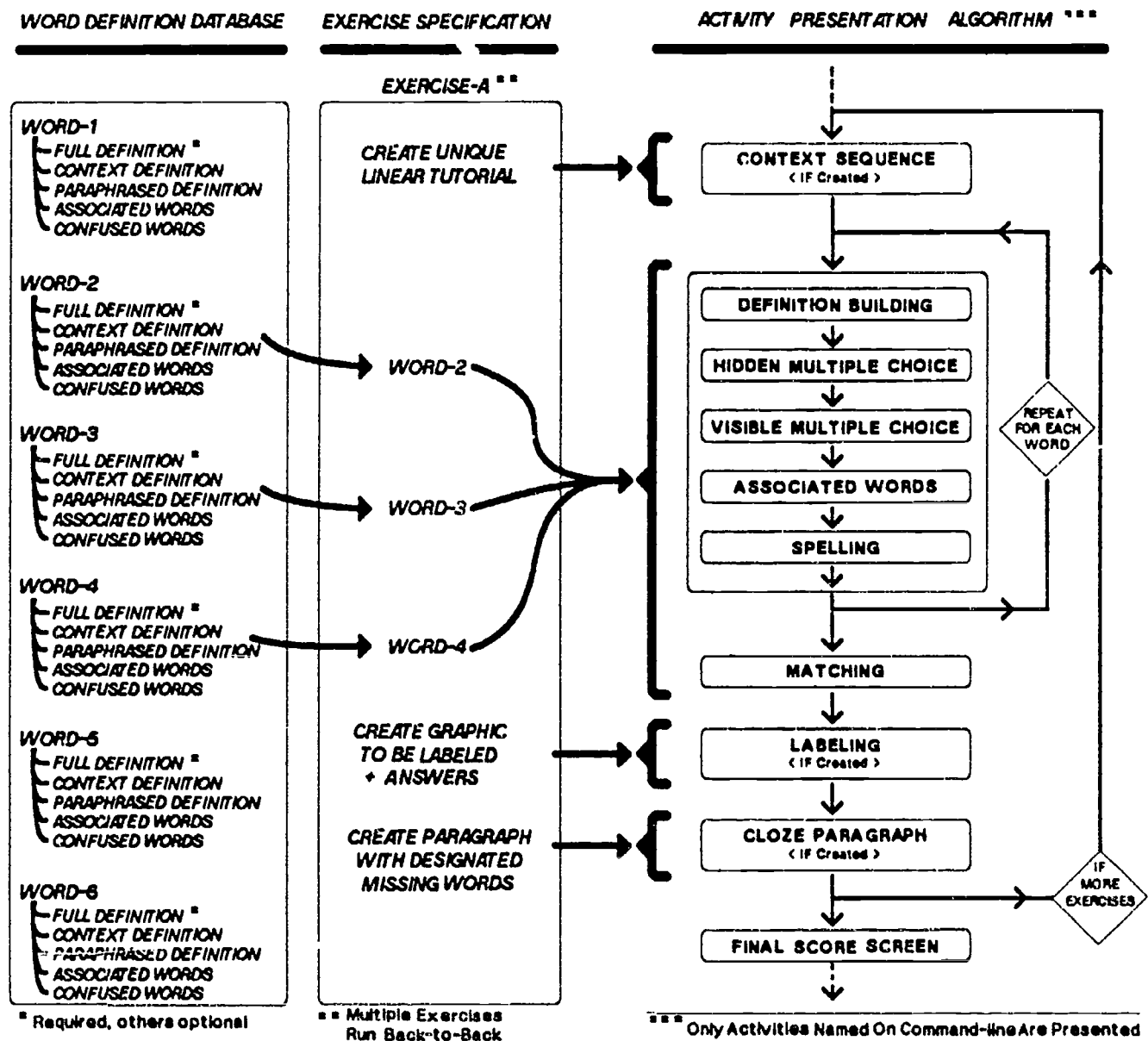


Figure 4. Vocabulary word database, exercise specification & presentation algorithm.

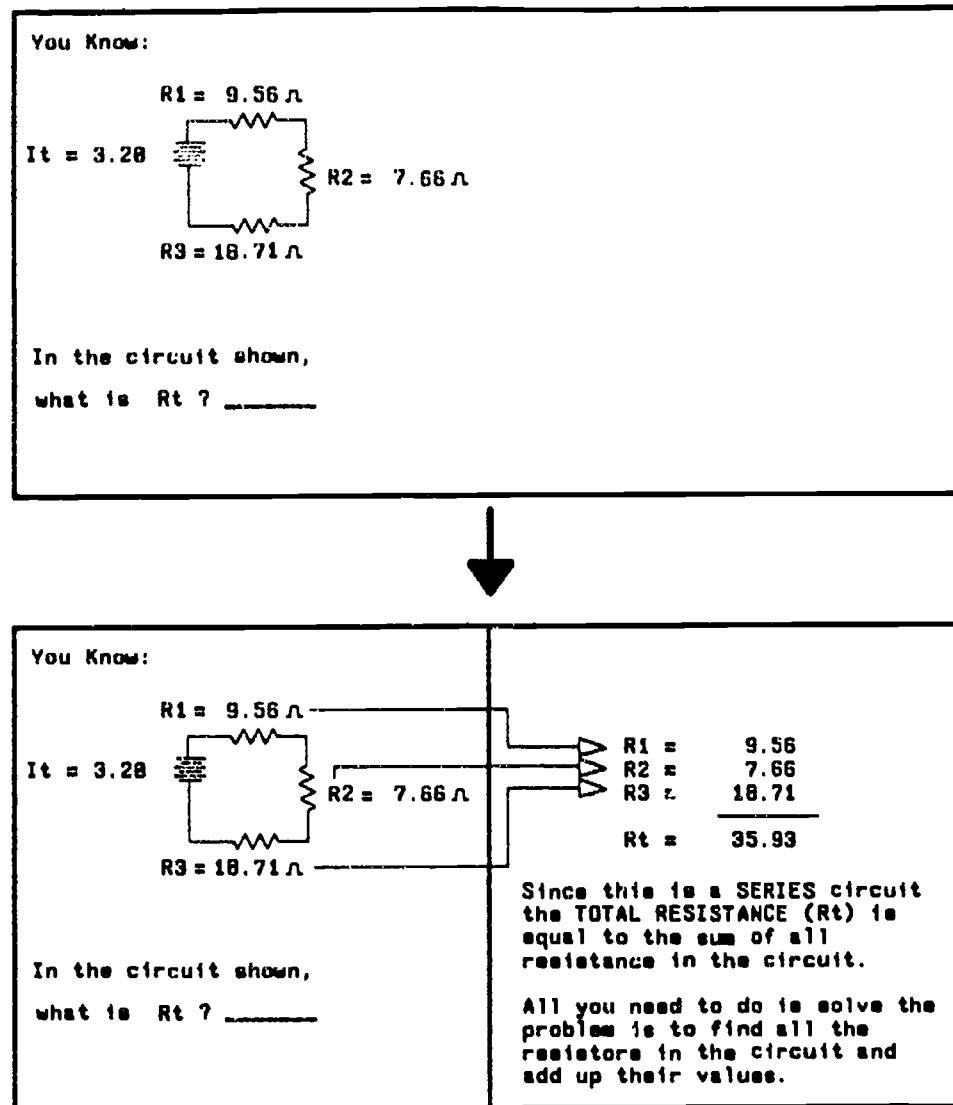


Figure 5. Circuit practice with numerical instances (top screen) and feedback (bottom) varying from problem to problem.

(3) **Free Form Templates:** A third use of templates allows unique instructional strategies to be specifically arranged by a developer in a hybrid generative CBI technique. This program is essentially a general CBI authoring package in which the 'dynamic placeholder' variable slots have been provided at strategic points. The program provides ten question templates for common circumstances such as receiving student answers from various multiple choice menus, pointing at locations on the screen, typing in answers, etc. These question templates have associated presentation, answer analysis and feedback options in which the 'dynamic placeholder variable slots are allowed. The generative nature of this application is enabled by a higher level control language which exports variables to the dynamic placeholders in the question templates. The language is used to name the specific templates to be presented, and permits calculations, manipulation of variables, and looping schemes. The question template can then import these variables into the slots of textual displays shown to students as windows or multiple choice menus, in answer analysis category slots associated with potential answers, and in the displays associated with the feedback resulting from each answer category.

This package allows the author to create applications where a large number of practice problems are given and where each problem is unique from trial to trial. Figure 5 illustrates an instance of this technique with a calculation problem on a series circuit that used different numeric values from problem to problem. This example involved randomly generating numbers in a specified range, exporting those numbers to dynamic placeholder slots in a single question template containing a graphic of the circuit, calculating the correct answer, exporting the correct answer to the answer analysis slot, and exporting components of the correct answer for incorrect answer feedback in the form of a worked out problem. This application used several different types of calculation problems, each with an appropriate graphic. Other examples using this technique have also been practice applications, with the simpler examples being practice on sequences of semaphore letters and recognition of various symbols. More complex examples vary the type of problem and present feedback working out the problem steps with variables comprising the intermediate calculations of the problem. Examples of these have been with other circuits, resistor color codes, and with arithmetic using fractions and signed numbers.

This package generalizes the use of templates so they can be used over and over again for sophisticated generative CBI lessons. Developers are allowed lower level run-time control over the the contents of the 'dynamic placeholders', control over the details of the interface, and process control in the form of a miniature programming language. The cost of this level of control and added complexity is that authors must have greater levels of experience.

Table 2. Degree of Standardization

Name of Application	Student Interface	Ques-Ans Database	Tutorial	Feedback	Process Control
SEMNET	Standard	Standard	None/ Minimal	Standard	Strict Algorithm
VOCAB	Standard	Standard	Preface	Minor Variability	Some Control
FreeForm	Unique	Unique	Unique	Unique	Unique

### Discussion

The three generative CBI applications discussed vary in the degree to which they standardize some of the factors cited earlier in Table 1. Table 2 summarizes their degree of routinization. The semantic network package standardizes all of the factors shown in specializing for repetitive quizzing on facts and images. A minimal tutorial is presented by one program only in the sense that the facts to be quizzed are listed just prior to the quiz. The technical vocabulary program also standardizes many of the features, but provides a unique tutorial only as a preface to the testing/learning activities. Certain of those activities provide feedback that varies depending upon entry of word database tags for designated confused words or a general feedback for other wrong responses. Some degree of process control is available in the sense that activities may be selected in various combinations by the author. The final package provided free form templates with slots in which any prearranged content could be inserted for the question, answer analysis and feedback. The price for this flexibility was a greater authoring overhead and standardization was offered only in the sense that a selection of predefined question templates was available.



The three packages discussed above illustrate generative CBI techniques since they involve new instances being generated from components not previously assembled in their complete finally delivered form. That is, the programs produce output determined at run-time rather than simply presenting completely elaborated previous screens. All of these programs are reusable for new instructional content and they vary to the extent that the content is cast simply as a database for a standard student interface or is configurable in unique presentations. While there are variations in the complexity of such themes, these techniques are an intermediate step between conventional frame based instruction and more sophisticated artificial intelligence techniques (cf. Kearsley, 1987). By contrast, generative CBI is generally easier to develop and may be a more manageable technique for authors with low levels of programming expertise.

A general limitation of generative or template techniques is that they depend upon some degree of presentation similarity from instance to instance. Thus, the beneficial factors favoring routinization cited above are also liabilities when more flexibility is desired in tailoring the interface, the form of the questions, associated feedback and exercising process control. The design challenge is to specialize and yet leave flexibility via having analyzed the instructional situations sufficiently to cover the demands created by varying instructional contents and user preferences. With novice authors one hazard of generative CBI, or of using templates, is that only portions of the finished presentation may be seen during its creation. Principles of interface design might argue that the discrepancy between authoring and conceptualizing the final output to students would reduce the "directness" and ease of the process (cf., Hutchins, et. al., 1986). Different instances of generative CBI vary in this regard and the availability of a built in "tryout" function can be used to more immediately show the final state during authoring. Thus, while the automaticity, compactness and power offered by generative CBI can ease development by avoiding laborious creation of many screens, those factors are pitted against the immediacy of seeing the final product to be delivered. Similar general comments can also be made about the trade-off between the number of program options and the ease and rapidity of developing CBI and their interaction with the users' initial experience level and learning curves.

In using previously developed generative CBI packages, an initial analysis on the part of the user is essential to determine if the package will fulfill the intended need. Envisioning the final product with the selected application may be difficult for some types of instruction. Developing a trivial example with the package may aid in seeing if it has all the power desired for the intended application. Such determinations should still be seen as subordinate parts of the total decision of whether to computerize the instruction at all, or decisions about what portions of the instruction will yield beneficial investments from the additional effort required to develop CBI. Thus, in addition to *how* to develop the instruction, a preliminary determination must be made as to *when* CBI is appropriate, and *what* should be computerized. CBI might be better thought of as a part rather than a whole for many training courses since there may be no point in automation unless some benefit to CBI can be identified. Rather than computerizing entire curricula, a more rational path may be to identify selected CBI applications that offer an improvement for specific training objectives when integrated with conventional instruction. Some practical reasons for using CBI might be the offering of a learning capability not possible with conventional methods, reduced costs compared to more high fidelity trainers, supplementing instructor resources by automating routine instructional objectives, standardizing instruction over many sites, and time savings owed to individualized instruction (cf., Wetzel, et. al., 1987).

#### References

- Hutchins, E.L., Hollan, J.D., & Norman, D.A. (1986). Direct manipulation interfaces. In D.A. Norman & S.W. Draper (Eds.) *User Centered System Design* (pp. 87-124). Hillsdale, NJ: Lawrence Erlbaum.
- Kearsley, G. (Ed.) (1987). *Artificial intelligence and instruction: Applications and methods* Reading, Mass.: Addison-Wesley.
- Wetzel, C.D., Van Kekerix, D.L., & Wulfeck, W.H. (1987). *Analysis of Navy technical school training objectives for microcomputer based training systems*. Technical Report NPRDC TR 88-3 (San Diego, CA: Navy Personnel Research and Development Center).

# ANALYSIS OF HUMAN BRAIN ELECTRICAL ACTIVITY: TOWARDS REAL-TIME PREDICTION OF HUMAN PERFORMANCE<sup>1</sup>

Leonard J. Trejo

Neurosciences Division, Navy Personnel Research and Development Center  
San Diego, CA 92152-6800

## INTRODUCTION

Current research in the Neurosciences Division of the Navy Personnel Research and Development Center (NPRDC) is driven by the Navy's need for better methods of assessing the performance of combat system operators, particularly for predicting the ability of operators to continue to make accurate decisions under heavy workloads. The demands of modern combat systems have the potential for exceeding the capacity of the human to accurately process information, especially during times of great stress. The capacity of the human to perceive, integrate, remember, and use information may be challenged when the individual is monitoring radar and sonar displays, operating electronic warfare systems, or flying aircraft. Exceeding the capacity of the human operator in such situations may impair decision making and could result in costly tactical errors.

Although much is being done to improve the hardware reliability of combat systems, not enough is being done to improve the performance of system operators. The most unpredictable element in combat systems is often the human operator. Traditional personnel testing and training technologies have not eliminated this unpredictability. In part, this is because traditional methods tend to measure or enhance what a person knows rather than how a person processes information.

In this paper, I summarize experiments in which event-related potentials, or ERPs, were examined as potential predictors or correlates of decision-making performance of combat system operators. [ERPs are small electrical waves which are recorded from electrodes on the scalp. Unlike the electroencephalogram, or EEG, ERPs are always synchronized with an environmental event, such as a visual or auditory stimulus.] I first describe relationships between individual measures of ERP amplitude, and performance of an air defense radar simulation in 30 military subjects. Next, I discuss the implications of the findings for real-time performance monitoring and performance enhancement. Finally, I describe ongoing studies of ERP measures and performance on other tasks, including signal detection, short-term memory and complex decision making. In this context, I also discuss some observations on individual differences in ERPs and event-related magnetic fields and their possible relevance to the selection and training of personnel.

In the air defense radar simulation task, AIRDEF, our approach was to demonstrate relationships between *first-order* and *second-order* measures of ERPs as correlates of task performance. First-order measures emphasize the central tendency within a group, or within an individual over a period. Such measures include the amplitude and latency of components in the average ERP, or average amplitude within intervals of the average ERP. Second-order measures emphasize changes in first-order measures across time or conditions. They can include either differences between first-order measures obtained under different conditions, or trial-to-trial variability of ERP amplitude within a condition. It has been hypothesized that first-order measures relate to the quantity of mental resources available for task performance, whereas second-order measures reflect the ability to shift resources from one task to another, or to resist distraction by focusing resources on a single task (Trejo, 1986; Trejo, Lewis, and Blankenship, 1987, 1990). We refer to the quantity of resources as *total capacity* and the ability to shift resources as *allocation range*. Figure 1, adapted from Defayolle, Dinand, and Gentil (1971), shows how first and second order measures of the ERP may correlate with task performance in high- and low-performing individuals.

## METHODS

We presented irrelevant visual stimuli (also called probes) to 30 male volunteers during a passive baseline period and during their performance of AIRDEF. Each subject performed AIRDEF at two levels

<sup>1</sup>The opinions expressed here are those of the authors, are unofficial, and do not necessarily reflect the views of the Navy Department. Approved for public release; distribution is unlimited.

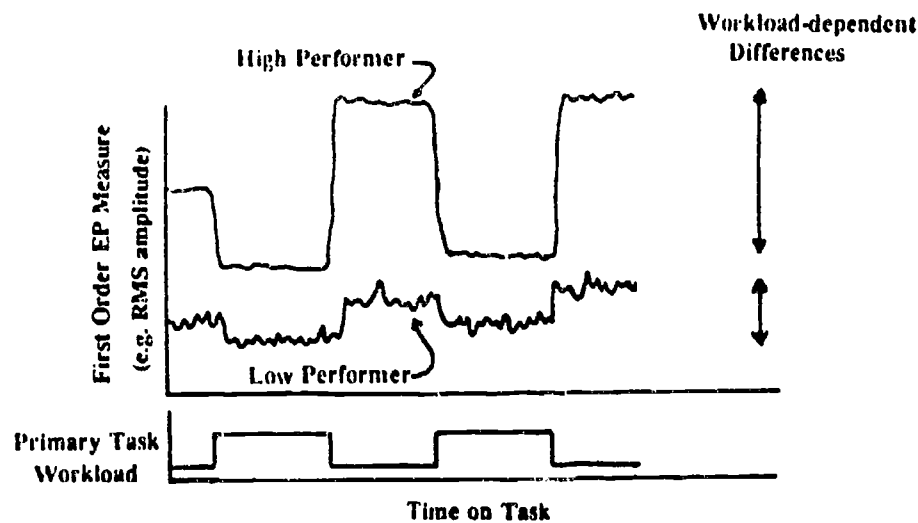


Figure 1. Hypothetical curves showing the relationship of first-order measures to performance on a complex task which is varying in workload. High performers are characterized by larger first-order mean values and greater allocation range. Low performers are characterized by lower first-order mean values and smaller allocation range. Adapted from Defayolle, Dinand, and Gentil (1971).

of workload, which were defined in terms of the rate at which targets appeared on the radar display. In level 1, 18 targets appeared over a four-minute engagement. In level 2, 36 targets appeared over a four-minute engagement. The probe stimuli were diffuse, low-intensity flashes of light with a duration of 16 milliseconds, presented at irregular intervals. These flashes filled the background of the same 13-inch color monitor used by AIRDEF, but had a negligible effect on the visibility of the AIRDEF display.

Figure 2 is a schematic diagram of the AIRDEF display. Subjects were required to detect hostile enemy targets, indicated by tracking numbers and radar "blips," and to fire weapons to kill these targets as close as possible to the maximum weapons range. Incoming targets varied in speed, and several targets could appear on the screen simultaneously. Targets that were not killed on time would reach the ship, resulting in a "hit." Performance was gauged by an overall skill rating that positively weighted kills according to their distance from the ship, and negatively weighted hits.

Under each condition, ERPs were recorded from electrodes covering the left and right frontal, temporal, parietal, and occipital areas of the scalp. A vertex electrode was the reference for all recordings. Each single ERP was first filtered (3 dB bandwidth 0.1-100 Hz), then sampled at 256 Hz, digitized, and stored by a computer. Signal-average waveforms were computed from six artifact-free ERPs for each condition. Each point in the signal-average waveform was the time-indexed average of the six single ERPs. These waveforms were then digitally filtered (0.5-25 Hz) and divided into eight adjacent, non-overlapping time windows, approximately 50 milliseconds wide, that spanned the range between 50 and 450 milliseconds after stimulus onset. The root-mean-square amplitude (RMS) of the waveform was computed in each window of the waveform. For brevity, we will refer to this measure as the RMS-a. This RMS-a value was used as the dependent variable in a repeated measures analysis of variance. Within-subjects factors were electrode position and window latency (time after stimulus onset).

## RESULTS

Figure 3 shows average ERP data for one subject as a function of recording sites and conditions. In this subject, as in most subjects, the RMS-a of the ERP decreased as a function of workload on the AIRDEF task, being largest in the baseline condition, smaller in level 1, and smallest in level 2. We analyzed this decrease in RMS-a as a function of recording sites and time after stimulus onset, comparing the ERP in the active AIRDEF engagements to that in the baseline condition. Across subjects, five site-window combinations of the ERP showed significant reductions of RMS-a with workload, ranging between 29.4 percent and 45.8 percent. We focused on these workload-sensitive site-window combinations in the following analyses of correlations between ERPs and task performance.

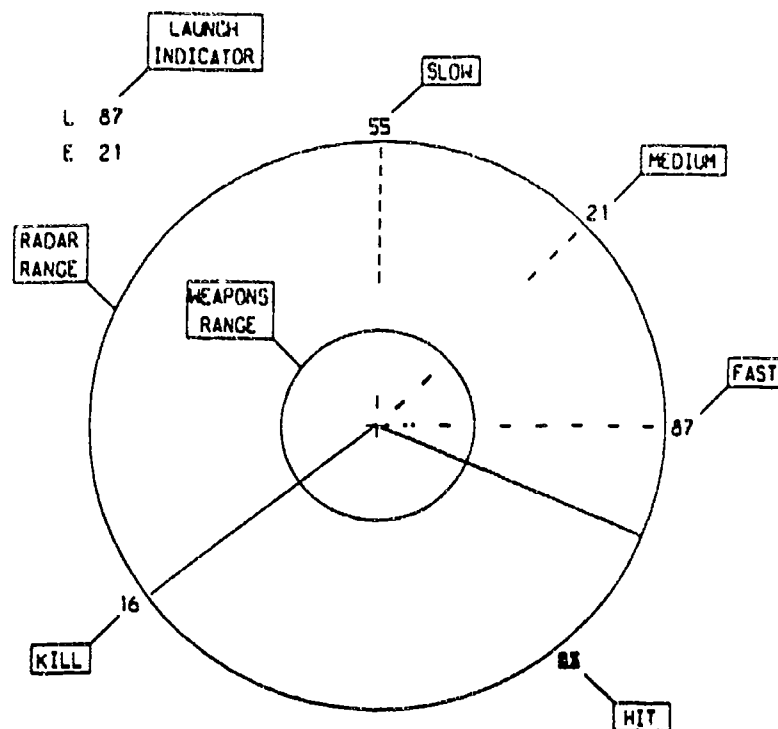


Figure 2. Schematic diagram of the air defense radar simulation. The cross in the center shows the location of the subject's ship and the inner circle marks the maximum weapons range. The outer circle marks the maximum detection range of the radar. Numbers along the outer circle identify incoming targets. Slow, medium, and fast targets show up as "blips" spaced at different intervals. Firing of weapons may result in kills, actually marked by solid green lines. Failure to fire results in hits, actually marked by solid orange lines.

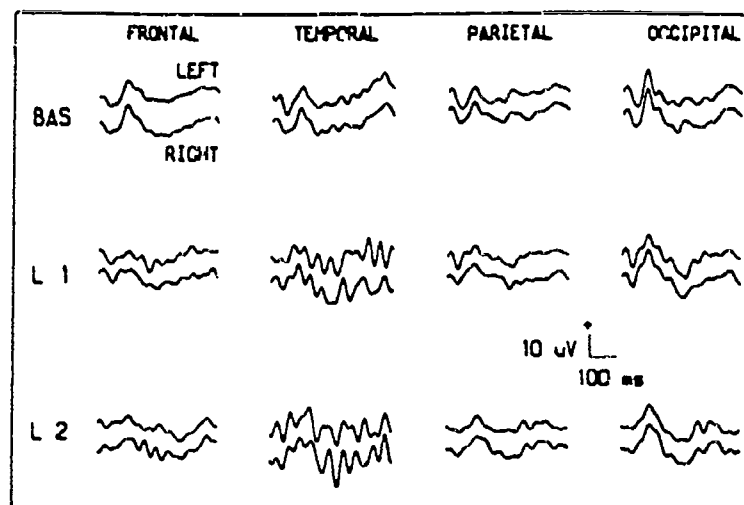


Figure 3. Average evoked potentials of one subject are shown for eight recording sites (left/right frontal, temporal, parietal, occipital) and across three experimental conditions (baseline, level 1, and level 2). Traces begin at the time of stimulus onset and extend for 500 milliseconds. Calibration bars represent 10 microvolts and 100 milliseconds. Recordings were referenced to a central midline electrode.

To test the predictive value of first-order measures of ERP amplitude, we examined the correlation between the five site-window RMS-a values in the baseline condition and subsequent performance in the active AIRDEF engagements. Table 1 lists the five workload-sensitive site-windows of the ERP, and shows their correlation with basic and global measures (hits, kills, and skill rating) of AIRDEF performance. A clear trend towards positive correlations between good performance and RMS-a in the five site-windows was observed. RMS-a in frontal window 6, with a latency of 330 milliseconds after stimulus onset, had the highest and most consistent correlations with performance in levels 1 and 2.

**Table 1. Correlations of First-order Measures and AIRDEF Performance**

*A. Level 1 AIRDEF Performance*

<i>Site-window<sup>1</sup></i>	<i>Kills</i>	<i>Hits</i>	<i>Skill rating</i>
Frontal-W2	0.23	-0.16	0.32*
Frontal-W5	0.27	-0.31*	0.32*
Frontal-W6	0.30	-0.29	0.45**
Parietal-W4	0.35*	-0.31*	0.40*
Occipital-W4	0.28	-0.23	0.34*

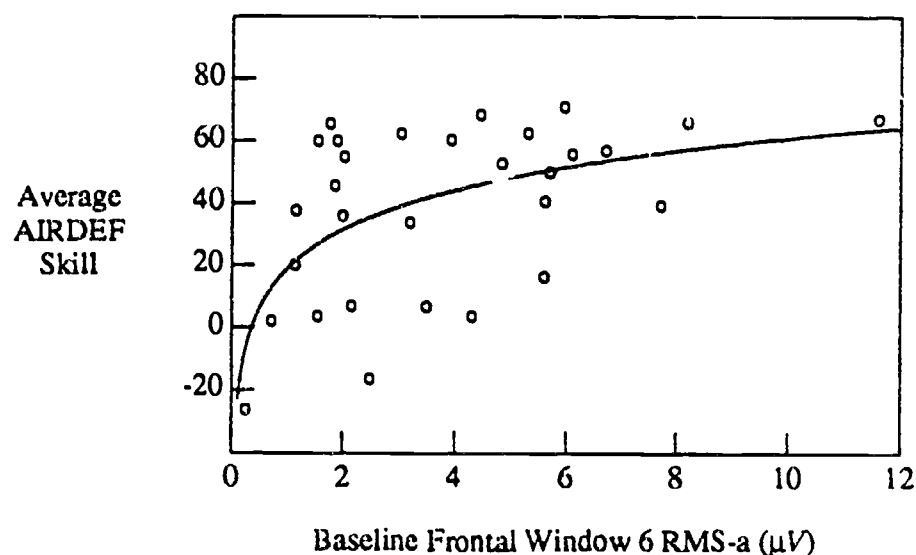
*B. Level 2 AIRDEF Performance*

<i>Site-window<sup>1</sup></i>	<i>Kills</i>	<i>Hits</i>	<i>Skill rating</i>
Frontal-W2	0.25	-0.35*	0.31*
Frontal-W5	0.17	-0.22	0.18
Frontal-W6	0.42**	-0.42**	0.38*
Parietal-W4	0.21	-0.24	0.14
Occipital-W4	0.10	-0.16	0.09

<sup>1</sup> Average of RMS-a values for homologous sites in both hemispheres.

\* One-tailed test,  $p < .05$ ; \*\*  $p < .01$ .

We further analyzed frontal window 6 by performing a non-linear least squares regression of average overall AIRDEF performance as reflected by the skill rating. The results are shown in Figure 4. A logarithmic function predicted 30 percent of the variance in AIRDEF performance using baseline values of frontal window 6 RMS-a in our sample of 30 subjects.



**Figure 4.** Shown are the 30 subjects' paired values of RMS-a in baseline frontal Window 6 and average skill across AIRDEF 18 and 36 target conditions (levels 1 and 2, respectively). A non-linear regression was significant, indicating the presence of a predictive relationship between the first-order measure and average AIRDEF performance ( $y = 42.1 \log x + 18.94$ ,  $F_{1,28} = 12.05$ ,  $p < .0017$ ,  $r^2 = .30$ ).



Next, we examined the value of second order ERP measures as predictors of AIRDEF performance. For each subject, in each of the five workload-sensitive windows listed in Table 1, we measured the change in RMS-a from baseline to active AIRDEF engagement for each subject. Since most subjects showed a decrease in RMS-a in active AIRDEF engagements as compared to the baseline condition, the following analysis focuses on the magnitude of these decreases. Specifically, we looked at the correlation of these decreases in RMS-a with both the global performance score (skill) and basic performance measures (hits and kills). The results are shown in Table 2.

A clear pattern of negative correlations was observed between performance measures and the decreases in ERP amplitude due to workload. Conversely, positive correlations were observed between error measures and ERP amplitude decreases. The most consistent ERP correlate of performance was frontal window 6, which was the average RMS-a over the frontal sites at a latency of 330 milliseconds.

We further analyzed frontal window 6 by computing a normalized difference score as the difference between active and baseline RMS-a values divided by their sum. Figure 5 shows the linear regression of average overall AIRDEF performance on this difference score in our sample of 30 subjects. The regression was significant, accounting for 27% of the variance in performance. Subjects who showed large reductions in frontal window 6 RMS-a tended to perform better on AIRDEF than those who showed an increase or no reduction.

## DISCUSSION

The results of this study are consistent with an information processing model in which neural responses to irrelevant probe stimuli, i.e., probe ERPs, predict and covary with human performance. We found that first-order measures of probe ERP amplitude (RMS-a) appear to index total capacity by significantly predicting subsequent AIRDEF task performance of military subjects. The pattern of correlations between the workload-sensitive first-order measures we examined and AIRDEF performance measures clearly support a total capacity hypothesis.

Because the first-order measures predicted AIRDEF performance in advance, it is possible that these measures reflect aspects of subjects' general information processing ability or intelligence. Although the data were not described here, we also found that a direct relationship held between first-order probe-ERP measures and on-job performance in the same 30 subjects (Trejo, et al. 1990). A high-performance group of subjects exhibited higher mean RMS-a values than a low-performance group in four of the five workload-sensitive windows. Two of these differences, frontal windows 2 and 6, corresponding to post-stimulus latencies of 227 and 330 ms, were significant.

Further evidence for a link between general ability and event-related brain activity has been observed in our laboratory with neuromagnetic measures. In a related study, for the same 30 subjects, we took neuromagnetic measures of the event-related magnetic field for visual checkerboard stimuli (Lewis, Trejo, Nunez, Weinberg, and Naitoh, 1987). In a condition similar to our AIRDEF baseline, event-related magnetic fields from the occipital region of the head were significantly larger in a high on-job performance group than in a lower-performance group.

In the present study, we found that second-order measures (changes in first-order measures) of probe-ERP amplitude were clearly related to task performance. These changes showed up as workload-related reductions in amplitude of the probe ERP at specific sites and in specific time windows after stimulus onset. There was a pattern of significant negative correlations between these amplitude reductions and AIRDEF task performance, which supported the allocation range hypothesis. The allocation range explanation for this is that subjects who were able to shift more resources to the task in the active conditions, as compared to baseline, performed better on the task. These shifts are reflected in the probe-ERPs as amplitude reductions. The proportion of task-performance variance accounted for using second-order measures was higher than that found for first-order measures.

As with the first order measures, other analyses we performed with second-order measures indicated a pattern of differences between high and low on-job performance groups. The differences we observed were in the same direction as those we found for task performance. A larger average decrease in probe-ERP amplitude between baseline and active engagement was observed for high performers than for low performers in four of the five site-window combinations tested, including frontal window 6. Thus there is evidence from second order measures that allocation of resources is predictive of general ability, such as



**Table 2. Correlations of Second-order Measures and AIRDEF Performance**

*A. Level 1 (18 target condition) AIRDEF Performance*

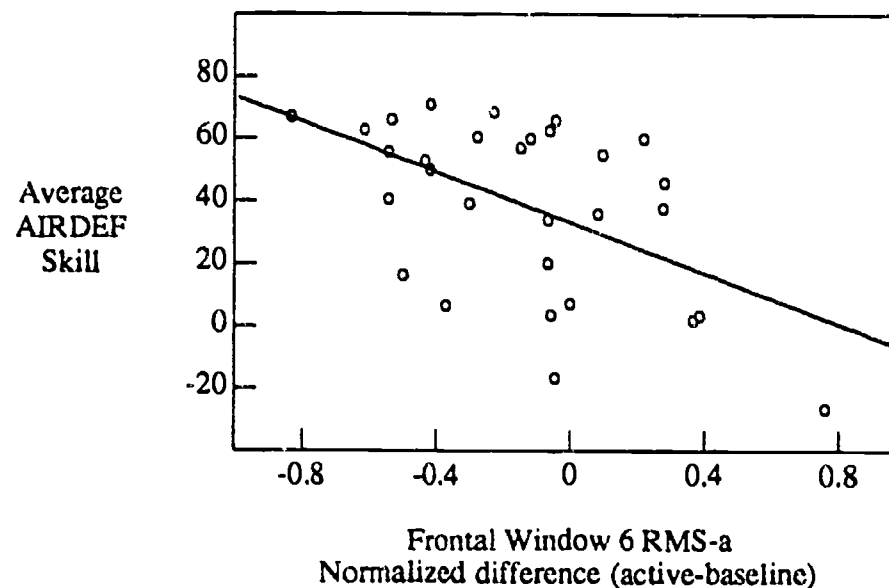
<i>Site-window<sup>1</sup></i>	<i>Kills</i>	<i>Hits</i>	<i>Skill rating</i>
Frontal-W2	-0.12	0.12	-0.10
Frontal-W5	-0.33	0.38*	-0.30
Frontal-W6	-0.47**	0.52**	-0.56***
Parietal-W4	-0.43**	0.43**	-0.40*
Occipital-W4	-0.47**	0.47**	-0.36*

*B. Level 2 (36 target condition) AIRDEF Performance*

<i>Site-window<sup>1</sup></i>	<i>Kills</i>	<i>Hits</i>	<i>Skill rating</i>
Frontal-W2	-0.16	0.29	-0.20
Frontal-W5	-0.07	0.20	-0.21
Frontal-W6	-0.44**	0.47**	-0.38*
Parietal-W4	-0.27	0.32	-0.15
Occipital-W4	0.01	0.13	0.00

<sup>1</sup> Average of RMS-a values for homologous sites in both hemispheres.

\* One-tailed  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .



**Figure 5.** Shown are the 30 subjects' paired values of the RMS-a normalized difference score for frontal Window 6 and average Skill across AIRDEF 18 and 36 target conditions (levels 1 and 2, respectively). A linear regression was significant, indicating the presence of a direct relationship between the second-order measure and average AIRDEF performance ( $y = 40.5x + 33.19$ ,  $F_{1,28} = 10.47$ ,  $p < .0031$ ,  $r^2 = .27$ ).

that measured by on-job performance criteria.

The most consistent relationships between performance and RMS-a, both first and second order, were found with frontal window 6. This window represents average amplitude in the probe-ERP recorded at frontal electrode sites referred to the vertex, or midline central electrode. The latency range of this

window is 305 to 355 ms. The primary component of the ERP that occupies this latency range is known as the P300. Many studies have shown relationships between P300 amplitude and cognitive processing (reviewed by Gopher and Donchin, 1986). Although P300 typically exhibits a maximum amplitude on the midline centro-parietal region, it may also be recorded at frontal sites. Thus, a voltage difference between frontal sites and vertex will probably reflect P300 amplitude.

## IMPLICATIONS

Our results demonstrate that ERP waveforms recorded for irrelevant probe stimuli, presented either before or during the performance of a complex decision-making task, provide information about the performance of an individual on that task. Because such probe stimuli are unobtrusive, it is possible to incorporate them into many tasks that require attention to visual displays without degrading task performance. By monitoring the brain's responses to these probes, information about the current performance state of the operator could be obtained in real-time. This information could be used to allow the system to sense operator overload or inattention, and to trigger interventions that could improve or sustain performance. Such interventions could be as simple as a warning light or buzzer, or as complex as adaptive decision-aiding procedures embedded in the task itself.

The correlation with performance of the probe-ERP data described here is too low to be of practical value in real-time monitoring. Three Navy laboratories are addressing this issue. At NPRDC, the research is aimed at isolating different levels of cognitive processing and examining the sensitivity of probe-ERP methods to variations in workload and to individual differences in ERP waveforms. The levels of processing are a) sensory-perceptual, b) short-term memory, and c) higher-level decision making. The approach being taken is to decompose the AIRDEF simulation into subtasks that emphasize one of these three levels of processing. Each subtask is performed several times by experienced combat system operators while probe-ERP measures are recorded. Both irrelevant and task-relevant probes are being examined. At the Naval Health Research Center, the research aims to develop probe-ERP indicators of sonar operator alertness. The N2 wave, an early component of the ERP elicited by irrelevant auditory probes, has shown potential for predicting lapses in performance a few minutes before they occur in an auditory signal detection task. At the Naval Aerospace Medical Research Laboratory, other research is examining the sensitivity of probe-ERP methods to the degradation in performance that occurs on variety of tasks performed during sustained and continuous operations. This information could be useful in improving work-rest cycles, or for predicting how well a given individual may tolerate assignment to continuous work without rest.

In all three laboratories, much effort is being devoted to improving signal processing techniques, which will improve the correlations that can be obtained between ERPs and performance. Although noise or variability in ERP data is high, several signal processing algorithms can improve the quality of the data. One promising approach is the application of neural networks to ERP signal processing and classification. ERP waveforms are highly idiosyncratic, being regular within an individual, but differing between individuals. Neural networks can be trained by example to recognize the features of an individual's ERP data that are most significant for predicting task performance. A research project at Temple University, sponsored by NPRDC, is examining such applications of neural networks. Initial results indicate that a neural network may correctly classify as many as 90 percent of single trial ERP waveforms as to their correspondence to a subsequent signal detection response. If such accuracy can be obtained in more complex tasks, it may be possible to store the parameters learned by neural networks for an individual and activate them when that individual is about to begin performing a task in a combat system.

Although the Navy research I have described here is directed towards problems experienced by combat systems operators, there are clear analogies between combat systems operators and operators of systems or equipment used in the private sector. A partial list of these includes air traffic controllers, airline pilots, truck drivers, and nuclear power plant operators. In all of these occupations there is the potential for operator overload, fatigue, boredom, and other errors which may be detected or averted by

probe-ERP methods.

## REFERENCES

- Defayolle, M., Dinand, J. P., & Gentil, M. T. (1971). Averaged evoked potentials in relation to attitude, mental load and intelligence. In W. T. Singleton, J. G. Fox, & D. Whitfield (Eds.), *Measurement of man at work*. New York: Van Nostrand Reinhold Company.
- Gopher, D., & Donchin, E. (1986). Workload—an examination of the concept. In K. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance*. Vol. II. New York: John Wiley.
- Lewis, G.W., Trejo, L.J., Nunez, P., Weinberg, H., & Naitoh, P. (1987). Evoked neuromagnetic fields: Implications for indexing performance. In K. Atsumi, T. Katila, M. Kotani, S. J. Williamson, & S. Ueno (Eds.), *Biomagnetism 1987, Proceedings of the 6th International Conference on Biomagnetism*. Tokyo: Tokyo Denki University Press, pp. 266-269.
- Trejo, L.J. (1986). *Brain activity during tactical decision-making: I. Hypotheses and experimental design* (NPRDC Tech. Note 71-86-6). San Diego: Navy Personnel Research and Development Center.
- Trejo, L. J., Lewis, G. W., & Blankenship, M. H. (1987). *Brain activity during tactical decision-making: II. Probe-evoked potentials and workload* (NPRDC Tech. Note 88-12). San Diego: Navy Personnel Research and Development Center.
- Trejo, L.J., Lewis, G.W., & Blankenship, M.H. (1990). *Brain activity during decision-making: III. Relationships between probe-evoked potentials, simulation performance, and on-job performance*. (NPRDC Tech. Note 90-9). San Diego: Navy Personnel Research and Development Center.

LEARNINGS FROM AN AFFIRMATIVE ACTION EFFORT:  
Minorities and Women as Decision Makers, and the  
Implementation of a Governing Board's Directive

by

David Lopez-Lee

School of Public Administration  
U N I V E R S I T Y   O F   S O U T H E R N   C A L I F O R N I A  
c 1990

## ABSTRACT

After 180 faculty members retired from the nine colleges of the Los Angeles Community College District (LACCD), an intense affirmative action effort to hire minorities and women was initiated as a consequence of a "directive" from the LACCD's governing board. Out of 128 faculty hires, 49.2 percent were minority and 54.7% were women. This provided the backdrop for this study. Using the data on the 128 faculty hires, four hypotheses were tested; one hypothesis was confirmed, i.e., campuses with presidents who are minority, hire more minorities than campuses with presidents who are non-minority. A questionnaire approach was also employed to assess the relative utility of the activities engaged in by the various campuses in their affirmative action endeavors. The campuses which were more successful in their affirmative action efforts, were those whose presidents had more hiring discretion. Although not strongly suggested by the questionnaire responses, it was concluded, given the confirmed hypothesis, that the inclusion of minorities and women in selection and screening panels is a useful affirmative action activity.

LEARNINGS FROM AN AFFIRMATIVE ACTION EFFORT:  
Minorities and Women as Decision Makers, and the  
Implementation of a Governing Board's Directive

by

David Lopez-Lee

In the Spring of 1989, as a consequence of a generous early retirement incentive program initiated by the governing Board of Trustees of the Los Angeles Community College District (LACCD), the LACCD found that it might have considerably more than the 100<sup>1</sup> to 125 retirements it expected (out of 1,800 tenured faculty). The LACCD's Board of Trustees, realizing that it might have to hire in excess of 100 faculty by the beginning of the coming fall semester, viewed this as an unusual opportunity to initiate a strong affirmative action program. The Trustees engaged in an action unprecedented in the District's history--in April of 1989, three Trustees (representing the unanimous sentiment of the full seven member board) personally addressed all nine of the LACCD's college campus presidents at a Chancellor's Cabinet meeting. At that meeting they underscored the intensity with which they wanted an affirmative action effort in its impending faculty recruitment effort.

As it turned out, 180 senior faculty members retired, and the District ended up hiring 128 new faculty--49.2 percent (i.e., 63) were ethnic minority and 54.7 percent (i.e., 70) were women. From a broad affirmative action point of view, this compared favorably with those who retired: 80 percent of the faculty retirees were white and 61 percent were male. However, when the data were viewed with respect to specific affirmative action goals the results were



slightly mixed: Black faculty representation which was previously 11.1 percent, moved up to 11.6 percent (23 were hired), which was the established goal of the district; Hispanic faculty representation which was 8.5 percent rose to about 9.5 percent (23 were hired) compared to an established goal of 20.8 percent; and Asian representation which was 5.5 percent rose to 5.7 percent (14 were hired)--the established goal was 7.5 percent. The representation for women which was 40.0 percent rose to 41.2 percent (70 were hired) compared to the established goal of 49.8 percent. Since one might erroneously minimize the extensive effort made in hiring Blacks and Asians, it is important to note that 16 Blacks and 11 Asians opted for early retirement; slightly greater percentage gains were made for Hispanics, because only 7 Hispanics exercised the early retirement option.<sup>2</sup>

In my judgement, there is much to be learned from such affirmative action endeavors. Indeed, it was believed that the variation in affirmative action performance between the various campuses provided this investigator with an unusual opportunity to engage in a two-fold endeavor: 1. to determine if having minorities and women included as decision makers in the hiring process (in this instance as campus presidents) increases the hiring of minorities and women being hired, and 2. to assess the relative utility of the activities engaged in by the various campuses in response to their governing board's "directive."

With respect to the first endeavor, one comprehensive review of the relevant literature states (within the personnel evaluation context): "that the more dimensions two or more people differ on within an interpersonal situation, the greater the distortion of interpersonal communications."<sup>3</sup> By obverse inference, one should

be able to predict that the greater the similarity between people within an interpersonal setting, the less the distortion (i.e., the "better" the evaluation or assessment). Thus, with respect to our first endeavor it seemed reasonable to forward the following two hypotheses:

1. Minority decision makers more positively evaluate minorities than do their non-minority counterparts.
2. Women serving as decision makers more positively evaluate women than do men in such positions.

Continuing with the same line of thought, since minorities and women have experienced similar forms of discrimination, one would reasonably expect a common ground of sensitivity between both groups. Accordingly, the following two additional hypotheses were formulated:

3. Minority decision makers more positively evaluate women than do their non-minority counterparts.
4. Women serving as decision makers more positively evaluate minorities than do men in such positions.

The foregoing hypotheses are also consistent with the findings from another context. The presence of a woman mayor, and women on city councils/assemblies, were identified by Saltzstein as<sup>4</sup> advantageous to the hiring of the disadvantaged by a city.

Warner and Steel found that the higher the percentage of women on the city council the higher the percentage of women employed as police officers; and although they found that the presence of a female as opposed to a male mayor resulted in no significant effect, the results were in the same direction as those found for the city council.<sup>5</sup> The most important point to be derived from

these studies is this: as removed from the initial recruitment and selection processes as they are, elected officials apparently do impact those processes. In the present setting, although the campus presidents are not elected officials, they too are somewhat removed from the recruitment and selection process, but unlike the elected officials they are typically involved in the final decision making processes.

In addressing this paper's second endeavor, that of assessing the relative utility of the activities engaged in by the various campuses in response to their governing board's "directive," it was believed useful to take close scrutiny of the following four areas:<sup>6</sup>

1. Recruitment activities,
2. The construction of screening/selection panels,
3. Procedures for forwarding hiring recommendations. and
4. Affirmative action training.

What immediately follows is the methodology employed to address this paper's two-fold endeavor.

#### METHODOLOGY

It was necessary to operationalize the four "abstract" hypotheses for testing within the decision making context at hand; doing so, the following four hypotheses were developed:

1. Campuses with presidents who are minority, hire more minorities than do campuses with presidents who are not minority.
2. Campuses with presidents who are women, hire more women than do campuses with presidents who are not women.

3. Campuses with presidents who are minority, hire more women than do campuses with presidents who are not minority.

4. Campuses with Presidents who are women, hire more minorities than do campuses with presidents who are not women.

Since the relevant data for testing these hypotheses were already available, only the manner in which the data were statistically analyzed are described below (under research design). Thus, most of what is provided in this section is the questionnaire methodology employed to address the four listed areas of the hiring process.

### Subjects

In this study all nine college campus presidents were sent a questionnaire. During the hiring time frame in question (April 1989 through September 1989), three of the nine campus presidents were minority and four were women.

### Procedure

A questionnaire was sent to each president with a cover letter from the Chancellor enlisting their cooperation. The text of that cover letter is contained in Exhibit I. These questionnaires and associated cover letters were sent out in March of 1990, with a request that they be returned within two weeks. All of the questionnaires were returned as requested.

### Questionnaire

The questionnaire was a straightforward extension of the earlier listed four areas of the campus hiring process. The text of that questionnaire is contained in Exhibit II (one "broad" question was added in the event an area may have been overlooked).

## Exhibit I. Cover Letter Text

Dear:

You may recall that last Spring (April 27, 1989), three Trustees addressed the Cabinet. Realizing that the District had an unusual opportunity to hire more than 100 new faculty, the three Trustees (representing the unanimous sentiment of the full Board) underscored the fact that they wanted an affirmative action recruitment effort that had teeth.

As you know, the District performed well in its response to the Board of Trustees' charge. Nevertheless, I trust you will agree that it is always fruitful to engage in self-assessment, particularly in such a sensitive area. By so doing, we may improve our efforts in the future.

In that regard, I would like your responses to the brief attached questionnaire. I look forward to receiving your responses. It would be appreciated if I had your responses by April 2, 1990.

Thank you very much.

## Exhibit II. Questionnaire Text

Please read all of the questions before responding. (Please feel free to complete your responses on additional sheets of paper.)

As a consequence of the Board of Trustees' charge:

1. Did you do anything differently in your recruitment activities? (Please explain.)

Please describe your previous recruitment efforts (other than what is done by the District Office).

2. Did you do anything differently in the construction of your screening/selection panels? (Please explain.)

Please describe the previous construction of your screening/selection panels.

3. Did you do anything differently in the manner in which hiring recommendations are forwarded to you? (Please explain.)

Please describe how hiring recommendations were previously forwarded to you.

4. Did you do anything differently in the training of persons involved in affirmative action activities? (Please explain.)

Please describe your previous training activities in affirmative action.

5. Did you do anything differently in any other areas? (Please explain.)

## Research Design

It was anticipated that the four operationalized hypotheses would be tested by arraying the data in 2 x 2 contingency tables and subjecting the data to chi-square analyses. The a priori level of statistical significance for these analyses was set at .05.

With respect to the questionnaire responses from the nine campus presidents, it was anticipated that their responses would be rank ordered in terms of their relative success in addressing the LACCD's affirmative action goals, thereby allowing one to assess the extent to which their stated activities corresponded with such success.

## RESULTS

As the reader may observe by an inspection of Table 1, the first hypothesis was confirmed, i.e., campuses with presidents who are minority, hire more minorities than do campus presidents who are not minority. The data arrayed in Table 2, while consistent with the second hypothesis (i.e, campuses with presidents who are women, hire more women than do campuses with presidents who are not women) did not produce statistically significant results. Similarly, while the data arrayed in Table 3 is consistent with the third hypothesis (i.e, campuses with presidents who are minority, hire more women than do campuses with presidents who are not minority), it too did not produce statistically significant results. However, the data arrayed (in Table 4) to test the fourth hypothesis (i.e., campuses with presidents who are women, hire more minorities than do campuses with presidents who are not women) though not statistically significant are consistent with a contrary hypothesis.



Table 1

	<u>Faculty Hires</u>	
	Non-Minority	Minority
Minority Pres.	15 (37.5%)	25 (62.5%)
Non-Minor. Pres.	50 (56.8%)	38 (43.2%)

Chi-square = 3.37, df=1  
p<.05, one-tailed test

Table 2

	<u>Faculty Hires</u>	
	Male	Female
Female Pres.	30 (44.1%)	38 (55.9%)
Male Pres.	28 (46.7%)	32 (53.3%)

Chi-square = 0.01,  
not significant

Table 3

	<u>Faculty Hires</u>	
	Male	Female
Minority Pres.	17 (42.5%)	23 (57.5%)
Non-Minor. Pres.	41 (46.6%)	47 (53.4%)

Chi-square = 0.06,  
not significant

Table 4

	<u>Faculty Hires</u>	
	Non-Minority	Minority
Female Pres.	38 (55.9%)	30 (44.1%)
Male Pres.	27 (45.0%)	33 (55.0%)

Chi-square = 0.06,  
not significant

EXHIBIT III provides a terse summary of each campus president's responses to the questionnaire. Deleted from the original responses given were answers which were: non-reponsive, belabored elaborations, and "good feeling" statements. Each campus president's responses are listed in order of the relative success of that campus in hiring minorities (parenthetically provided is that campus' ranking in the hiring of women). The reader should take note: since each of the last two campuses listed made only three hires, they are an insufficient base on which to make inferences. Thus, in evaluating this data, one should only focus on the first seven campuses listed. It should also be noted that the primary reason that the campuses were sequentially listed by the percentage of their hires who were minority (rather than by the percentage of hires who were women) was because it resulted in a more substantive variation in the hiring performances of the seven campuses--the percent of minorities hired by these campuses ranged from 38 to 89 percent, while the percent of women hired by them ranged from 42 to 72 percent. As one might expect, although not statistically significant, the correlation between both rankings was positive ( $\rho = .40, p > .05$ ). Thus, while one could just as easily have made the hiring of women as the primary basis for listing the campuses in Exhibit III, in the present instance the more substantive basis for listing them was their relatively "better" success in the hiring of minorities.

### Exhibit III. Questionnaire Response Summary

Campus	Recruitment Differences	Screening/Selection Panel Differences	Hiring Recommendation Differences	Affirmative Action Train'g Differences	Other Differences
1 (7.5)*	Used District pool and ethnic networks; same before	Made A.A. rep a voting member; not done before.	Selec. committee forwarded 3 applicants unrankd; same before. President interviewed all applicants; same before.	--	--
2 (2)	As before continued to use District pool.	--	Selec. committee forwarded 3 applicants rankd; same before. President interviewed all applicants; not done before.	--	--
3 (4)	District application packets made available on campus; made calls to universities & A.A. candidates; continued using District pool. Before; local universities contacted in some cases.	Depts made effort to include persons from underrepresented groups on selection committees; same before.	Selec. committee forwarded 2 applicants unrankd; same before. President didn't interview; discussed applicants with V.P. of academic affairs and sometimes with dept chair &/or A.A. rep; same before.	Since many positions were being filled, alternate A.A. reps were trained to serve on selection panels. A.A. rep met with selec. committees to discuss A.A. goals/policies/procedures; same before.	--
4 (7.5)	Visited universities to recruit; continued using District pool.	Open issues: giving vote to A.A. rep and to non-divisional member.	Selec. committee forwarded 2 applicants unrankd; same before. President interviewed both applicants; same before.	--	--
5 (5)	Made calls to universities; continued using District pool.	Effort made on selection panels to include member of an underrepresented group; not done before.	Selec. committee forwarded 2 applicants unrankd; rankd before. President sometimes interviewed applicants if it was believed that A.A. was not well served; same before.	Members of A.A. committee met monthly to review A.A. laws & regs.; before A.A. materials were simply sent to members.	Applicants answered structured questions--no clarification or prompts allowed. A.A. rep present during selection committee's decision-making after interview.
6 (1)	Placed advertisements in newspapers; circulated announcements in ethnic networks; sent mass mailings to target populations (4 yr institutions and professional orgs.)	--	Selec. committee forwarded 2 applicants ranked; same before. President's decision based on inputs from V.P. of academic affairs & A.A. rep (only V.P. interviews applicants); same before.	A.A. video shown to all dept heads & administrators; president discussed A.A. goals; A.A. rep and president discussed A.A. issues; previously A.A. issues were only discussed by A.A. committee.	Closer monitoring of selection process. Larger number of committee recommendations set aside.
7 (9)	Professional associations & 4 yr institutions were contacted and sent materials.	--	Selec. committee forwarded 2 applicants unrankd; same before. President interviewed both applicants; same before.	A.A. video shown to all dept chairs & their A.A. plans discussed; not done before.	Some depts did no hiring this year & focused instead on recruiting to enrich the District pool, and then hire this coming year.
8 (3)	Had no retirements. Thus, very few positions to recruit for; continued using District pool.	--	Selec. committee forwarded 2 applicants unrankd; same before. However, President discussed final applicants with committee chair & V.P. of academic affairs--their preferences became evident; same before. President interviewed both applicants; same before.	--	--
9 (1)	Minimally impacted by retirements; continued using District pool.	--	Selec. committee forwarded 3 applicants unrankd; rankd before. President interviewed all applicants with academic affairs V.P.	Provided train'g in addition to that provided by A.A. District Director.	--

\*The first number is the ranking of the campuses (from most to least successful) in hiring minorities; the number in parenthesis is the ranking in the hiring of women.

## DISCUSSION AND POLICY IMPLICATIONS

To be sure, how a college campus responds to a "directive" from its governing board is not wholly captured by assessing what its president does or does not do. Nonetheless, it is the President of a campus that is held accountable for any "directive" he or she is given, not his or her staff, not any department chair, and not any faculty committee. In these types of investigative endeavors, however, there is frequently a risk of the wrong inferences being made. Specifically, on the basis of the statistical analyses performed on the 2 by 2 contingency tables, one might erroneously conclude that some of the campuses did a poor job in addressing affirmative action goals. It should be noted that the "splits" between female and male presidents, and minority and non-minority presidents, are required by the statistical technique employed. Such "splits" tend to give the misleading impression that the bottom halves of these "splits" are all poor performers, and that the top halves are all good performers. It should be apparent from the data presented in these tables, that these "splits" did not correspond with the affirmative action performance of the various campuses. This is also true of Table 1, which did produce a statistically significant outcome--confirming the hypothesis that presidents who are minority hire more minorities than do presidents who are not minority.

Since the campus presidents were somewhat removed from the recruitment and selection of faculty, the confirmation of this operationalized hypothesis seemed remote. This finding underscores the point: as removed from the initial recruitment and

selection process as they are, campus presidents can have a substantive impact in the final hiring decision. Thus, it is recommended that research be conducted which either involves a much larger set of campuses, or which addresses the three "abstract" hypotheses within a context wherein the decision maker is less removed from the hiring process (e.g., investigate the effects of having minorities and women on faculty selection panels).

As mentioned previously the questionnaire responses from the various campuses (EXHIBIT III) are listed in order of their success in hiring minorities. The reader will recall that since the last two campuses listed made only three hires each, only the first seven campuses listed provide any substantive basis on which to make any inferences. It is with this perspective in mind that the ensuing discussion is framed.

While the differences in responses between the various campuses in their stated affirmative action endeavors appear subtle, aggregations of their responses are worthwhile noting. Two areas, screening/selection panel differences and differences in the forwarding of hiring recommendations, merit particular scrutiny.

With respect to the construction of screening/selection panels, four campus responses are related to the "abstract" hypothesis confirmed in this study (i.e., minority decision makers more positively evaluate minorities than do their non-minority counterparts). Two of the four responses were of direct relevance--two campuses reported that they made an effort to include persons from underrepresented groups on their selection panels. The other two responses were indirectly related--one campus stated that

voting rights were given to the affirmative action representative on the selection panel (this person is often a minority or a women), and the remaining campus indicated that giving voting rights to the affirmative action officer was still an open issue (it is arguable whether this response should be considered consistent with the above hypothesis). In a vein that is consistent with the "abstract" hypothesis being addressed, rather than encouraging or making an effort to include minorities and women on such selection panels, it makes much more sense to require their inclusion. If a given department has no one from the various underrepresented categories, they could find and enlist such persons from: another campus, a private firm, or a related discipline. (This latter type of inclusion is apparently what the fourth listed campus meant in its statement that "giving voting rights to a non-divisional member on a selection panel is still an open issue."<sup>7</sup>) It should be noted that as early as 1975, the Carnegie Council on Policy Studies in Higher Education, in a comprehensive survey of affirmative action policies of colleges and universities (including community colleges) across the United States, found that many institutions required that faculty search and/or review committees<sup>8</sup> include minorities and women.

Campus responses having to do with differences in the forwarding of hiring recommendations to campus presidents resulted in a difference that while apparently trivial, may prove to be a solid foundation upon which to build effective affirmative action programs. Very simply, while the first two listed campuses reported that their selection committees recommended three applicants to the



president, the next five campuses reported that their selection committees recommended two applicants. Given that the performance of the first two listed campuses in meeting affirmative action goals was substantively superior to that of the other campuses, it is important to note that this numerical difference in recommended applicants was the only substantive distinction separating these two campuses from the others. Because of the ease with which this numerical difference may be dismissed, it is believed important that the implications of this numerical difference be viewed from a broader perspective. This difference, in effect, increases a decision makers discretion by one-third--a considerable amount. While we may be increasing a decision maker's "degrees of freedom" from only 2 to 3 choices, such an increase becomes clearly substantial when such decisions are replicated 20 to 30 times. (Whether such recommendations were ranked or not did not seem important, since, in one way or another, many of the members of such committees let their preference be known to the president.)

Campus efforts via their responses in recruitment differences, affirmative action training differences, and "other" differences, while interesting (and perhaps in some instances useful), did not appear to be associated with affirmative action "success." (It should be noted that all of the campuses select their faculty applicants from a common district pool, and that brief affirmative action training sessions were provided each campus by the district's affirmative action director.) If anything, it would appear that a greater amount of "activity and verbage" was inversely associated with affirmative action success.

There are two policy implications that readily spring forth from this study:

1. College institutions should require the inclusion of minorities and women on screening and/or selection committees (I see no reason why this requirement should not be expanded to include promotion and tenure review committees.)
2. College institutions, particularly those which are not successful in their affirmative action hiring endeavors, should expand their president's "degrees of freedom" in the hiring process, i.e., if he or she is presently forwarded two choices in the hiring process, expand it to three, and if he or she is presently being given three choices, expand it to four.

## FOOTNOTES

1. The Los Angeles Community College District serves all of the the City of Los Angeles and a considerable portion of Los Angeles County--there are over 4.5 million people within its boundaries. The nine campuses have an enrollment of about 108,000 students--about two-thirds of its students are ethnic minority or foreign-born.
2. Affirmative Action Program: For Faculty and Staff Diversity, Affirmative Action Programs, Office of the Chancellor, 1989-1990. The affirmative action goals in this report were specified for six departmental areas and had the following spreads: for Blacks the goals ranged from 10.9 to 11.9 percent, for Hispanics the goals ranged from 20.6 percent to 21.2 percent, and for Asians the goals ranged from 4.0 to 8.4 percent; the median values for these three groups were 11.6 percent, 20.8 percent and 7.5 percent respectively--these are the values used in this paper. The recent hiring data were obtained from the Office of Affirmative Action Programs of the Los Angeles Community College District.
3. D. Lopez-Lee, "Organizational Representativeness: An Operational Imperative for the Personnel Function," Public Personnel Management, Volume 8, Number 5 (1979), pp. 287-293.
4. G. H. Saltzstein, "Female Mayors and Women in Municipal Jobs," American Journal of Political Science, Volume 30 (1986), pp. 140-164.
5. R. L. Warner and B. S. Steel, "Affirmative Action in Times of Fiscal Stress and Changing Value priorities: The Case of Women in Policing," Public Personnel Management, Volume 18, Number 3 (1989), pp. 291-309.

6. These four areas of the hiring process are taken from: A Report of the Carnegie Council on Policy Studies in Higher Education, Making Affirmative Action Work in Higher Education (San Francisco: Jossey-Bass Publishers, 1975).
7. Ibid.
8. According to the Personnel Guide, B 473, Office of Human Resources, Los Angeles Community College District, August 22, 1979, a selection committee consists of the department head and a minimum of two faculty members elected by the appropriate discipline; if there is not a sufficient number of faculty members in the discipline to participate on the selection committee, the additional members should be elected from a related discipline.

Role of the Buros Institute of Mental Measurements as an  
Information Provider to Personnel Selection Specialists

Barbara S. Plake Director  
Buros Institute of Mental Measurements  
University of Nebraska-Lincoln

Paper presented at the annual meeting of the International Personnel  
Management Association Assessment Council Conference, June,  
1990, San Diego.

Role of the Buros Institute of Mental Measurements as an  
Information Provider to Personnel Selection Specialists

(Abstract)

The Buros Institute, which publishes the Mental Measurements Yearbook and Tests in Print Series, has the capability of serving as a useful resource to personnel selection specialists. The goal of this presentation is to increase personnel selection specialists' awareness of the Buros Institute's products and programs, especially as they relate to personnel selection specialists' professional needs.



# Role of the Buros Institute of Mental Measurements as an Information Provider to Personnel Selection Specialists

## Introduction

Instrumentation plays a major role in the job demands and decision information systems of personnel selection specialists. Often a personnel selection specialist has to identify a test that will provide information for a personnel-related decision. Basically, the personnel selection specialist has two options: (a) adopt an already developed test or (b) develop a test specifically for this assessment need. While it might be argued that every assessment situation is unique and therefore already developed tests are invalid for the particular assessment problem, there are several reasons why already developed instrumentation may be appropriate for the assessment need. First, there is research and empirical evidence to suggest that assessment needs are generalizable across context. Therefore, the unique dimensions of the assessment situation may not create an invalidity for the test use. Second, test development is very costly, time consuming, and demanding. Worse yet, the resulting test may not demonstrate the requisite psychometric properties needed for the assessment decision. Third, in applications where litigation may result from decisions based in part on the results of instrumentation, an instrument developed by

an independent, professional testing corporation may be perceived as more objective and defensible.

Therefore, an efficient strategy for identifying a test may be to locate a commercially available instrument with proven psychometric properties for the purposes of the instrument's application. However, this strategy of locating the appropriate commercially available test is not a simple process since there are several thousand commercially available tests on the market. The personnel selection specialist would need to devote a substantial number of hours just to develop a roster of tests that might be useful for the needed assessment application. Communicating with the test publishers to find out test availability and utility is of course desirable, however, the psychometric and personnel psychology backgrounds of sales representatives often leave a lot to be desired. Test companies also are not well known for their objectivity in advertisement or communication of test utility and performance. Thus, what is needed is a resource that provides retrieval information of test availability and critical analyses of the psychometric properties and utility of the commercially available instruments.

### Buros Institute of Mental Measurements

The Buros Institute of Mental Measurements was established in 1938 by Oscar K. Buros with just these types of needs in mind. The Institute publishes two products that, in combination, can be

extremely helpful in the identification of commercially available tests for possible use and evaluation of these instruments. The Tests in Print Series fulfills the first need by providing availability information on all in print, commercially available English language tests. This series is published periodically; TIP III appeared in 1983 and TIP IV is scheduled for publication in 1991. The Mental Measurements Yearbook Series began in 1938 and contains evaluations of commercially available tests written by professionals in the field. Coupled together, the TIP and MMY series serve as an important resource for the identification and evaluation of commercially available instruments.

#### Mental Measurements Yearbook Database

While the Mental Measurements Yearbook Series provides qualitative reviews of commercially available tests, there are limitations to the timeliness of any published volume in providing current accessibility and evaluative information about tests. This was particularly true of the earlier MMYs, which were published on a 6 - 8 year cycle. It was quite possible for a test to be issued by a publisher and go out of print between publications of the earlier MMYs. In an attempt to better serve the needs of the test community, the Institute established the Mental Measurements Yearbook Database in 1983. This database, vended by BRS Information Technologies and updated monthly, provides current access to test reviews as they become available for distribution. By using search algorithms, individuals accessing MMYD are able to

obtain combinational information involving test name, test classification (e.g., achievement batteries, language, neuropsychological, reading, intelligence, vocational, etc. based on Mental Measurements Yearbook classification schemes), test author, test publisher, publication dates, population served, scores indices, administration information, reliability and validity data, and price. Therefore, a personnel selection specialist could seek assessment instruments developed by a particular publisher to measure leadership style. The reviews contained in the Eighth, Ninth, and Tenth Mental Measurements Yearbooks, as well as some in process for the Eleventh, are now searchable through MMYD.

#### Publication Schedule

The Buros Institute has also adopted a more frequent publication schedule for the Mental Measurements Yearbook Series. Yearbooks are now published biennially, and a softbound Supplement is published in the years in between publication of the hardbound MMYs. This publications schedule was established to meet the needs of consumers who do not have ready access to MMYD yet who need timely access to test review information.

#### Other Institute Products and Programs

While the primary products of the Buros Institute are the IIP and MMY series, the mission of the Institute supports broadly based activities which further the improved development and use of tests

and test-related products. A number of additional products and programs have been developed to meet these broader goals.

Buros Library. On site with the Buros Institute is housed the Oscar K. Buros Library of Mental Measurements. This library contains Oscar's historic collection of texts and testing products in addition to the most comprehensive test collection in the world. This facility is available for use by professionals in the fields of educational, psychological, and industrial measurement.

Consultation Service. While not in the business of making recommendations of specific tests for specific uses, the Institute responds to questions regarding test availability and use from customers world-wide.

Buros Nebraska Symposium on Measurement and Testing. Each year the Buros Institute sponsors a symposium dealing with a critical issue in the measurement and testing fields. A list of previous symposium topics follows:

1. Social and Technical Issues in Testing
2. The Future of Testing
3. The Influence of Cognitive Psychology on Testing

4. The Computer as Adjunct to the Decision Making Process
5. Assessment of Teaching: Purposes, Practice, and Implications for the Profession
6. Curriculum-Based Assessment: Examining Old Problems, Evaluating New Solutions
7. Are Our School Teachers Adequately Trained in Measurement and Assessment Skills?

Future symposia still in the planning stages are focusing on (a) marriage and family assessment, (b) multicultural assessment, and (c) predictive validity for selection decisions. Subsequent to each symposium, a volume containing manuscript versions of presented papers and additional solicited chapters is published through Lawrence Erlbaum and Associates.

Applied Measurement in Education (AME). AME is a scholarly journal created to provide both a greater understanding of educational measurement issues and an improved use of measurement techniques in education. Its intended audience consists of both researchers and practitioners who are interested in research that has a likely impact on educational measurement practice. Sponsored by the Buros Institute, AME is published by Lawrence Erlbaum and Associates.



## Summary

The purpose of this paper was to provide personnel selection specialists with information about the Buros Institute of Mental Measurements. The primary products of the Buros Institute is the Mental Measurements Yearbook and Tests in Print series. The Mental Measurements Yearbook Database, offered through BRS Information Technologies, provides searchable access to completed reviews even before they are published in the Yearbook. The Buros Institute sponsors an annual symposium dealing with a critical issue in the measurement and testing fields. Through the Oscar K. Buros Library of Mental Measurements and professional consultation the Buros Institute also provides information to governmental agencies, public schools, and individuals. Further, in conjunction with Lawrence Erlbaum and Associates, the Buros Institute sponsors a scholarly journal dedicated to the application of educational and psychological measurement research to the educational process, Applied Measurement in Education.

Changes in the Buros Institute are aimed at broadening the scope of the measurement activities and at improving the current service to the measurement community. The Buros Institute is expanding its measurement-related activities in ways that aid the better use of tests and testing practices. By providing professional assistance, expertise, and information to consumers of commercially published tests, the Buros Institute hopes to foster meaningful and appropriate test selection, use and practice. Additionally, the Buros Institute hopes to encourage improved test development and measurement research through thoughtful, critical

analysis of measurement instruments and the promotion of an open dialogue regarding contemporary measurement issues.

## Hiring And Monitoring Consultants

by: Vicki Packman

The consultant you hire is a reflection on you as a professional and your department. A poor selection can haunt you for years. The goal then is to hire the very best consultant possible.

The first step in selecting a consultant is to determine exactly what you want them to accomplish, the time frame involved, and a budget.

The next step is to ensure that you have a large sample of candidates to choose from. This can be accomplished by surveying similar organizations/agencies, professional associates and coworkers for recommendations of consultants they have hired for similar projects.

The Request For Proposal (RFP) is the key to the selection process. This document should describe the project in explicit detail and should include the occupation codes and number of incumbents of the jobs involved, special requirements such as amount of time to be spent on site, and pertinent information such as labor agreements. The RFP should request a detailed methodology, proposed start and complete dates, proposed cost, information on the firm and its staff and references. The RFP should also state the criteria on which the proposals will be evaluated such as the experience of bidders on similar projects, experience and qualifications of staff, the consultant's history for completing projects on time and within budget, and the proposed completion time and costs.

The cover memo accompanying the RFP should list enclosures and specify the number of sealed proposals required and the recipient, the date and time due, consequences of late proposals, and whether or not contacting the recipient/requestor is allowed.

The contract is a legal bond between your agency and the consultant. A good contract should include the effective (from-to) dates; compensation parameters (coach travel only, hourly rate plus expenses only, reimbursement of travel time, and retentions); your agency's expectations such as receipts required, monthly progress reports, etc.; a cancellation clause; insurance requirements; establish responsibility for tax payments on fees and include a not-to-exceed dollar figure. Including a sample contract with your RFP may discourage some consultants from submitting a proposal. This is fine. A reliable consultant will not be discouraged by an explicit but fair contract.

The proposal is the consultant's opportunity to present themselves at their best and really impress you and your agency/organization. Therefore, consultants should be disqualified for being late; using poor grammar, spelling, etc.; presenting insufficient detail or inappropriate methodology; subcontracting part of your project; making mistakes such as omitting a page, referring to your agency by an incorrect name, etc., or for proposing costs or time which exceed your requirements and limitations.

Call the references supplied in the proposal. Believe it or not, consultants do provide the names of clients who are dissatisfied.

Consultants should be disqualified if references report they were consistently late, hard to work with, or over budget.

Consultants who, for whatever reason, fail to provide references should also be disqualified.

A review of proposals may determine two or more consultants that are equally qualified. These can be invited to make a presentation to members of your agency. Your final selection can then be based on their presentation, whether they answered your questions satisfactorily, how they interact with yourself and the people in your agency they will be working with and their overall presentation of themselves.

Monitoring/supervising the consultant hired involves a lot more time than most people expect. But your time, just like your money, is an investment you make towards a successful project. Never give a consultant carte blanche. Monitoring their work involves ensuring that the work adheres to their proposal, progress reports and billings are timely and accurate, and the people from your agency are pleased with the consultant's work. You are the client and you and your agency should be satisfied with the project and how it progresses.

Ask to receive the final report in draft form so you can make changes. Again, the final report is a reflection on you so you should check it thoroughly to ensure the methodology is sufficiently detailed, tables and footnotes are accurate, and there are no spelling errors or typos. Do not hesitate to request revisions . . . you are the client.

The final steps include meeting with everyone involved from your agency to obtain and document their approval of the project and deliverables, notifying the consultant the project has been approved, and paying the retention fees.



## The Applicability of the Situational Interview and the Patterned Behavior Description Interview for Entry and Managerial Level Jobs

Heidi H. Mrowal & Thung-Rung Lin, Los Angeles Unified School District,  
Los Angeles, CA

There has been a great deal of discussion over the past several years regarding the validity of the interview, and how that validity is increased if structured interviews are used, and further increased if these structured interviews are based on job analysis (Arvey & J. Campion, 1982; M. Campion, Pursell & Brown, 1988; Cronshaw & Wiesner, 1989; Harris, 1989; Wiesner & Cronshaw, 1988). Two methods, in particular, have received greatest attention: the Situational Interview (Latham, 1989; Latham & Saari, 1984; Latham, Saari, Pursell, & M. Campion, 1980) and the Patterned Behavior Description Interview (Janz, 1982, 1989; Janz, Hellervik & Gilmore, 1986). Both interview methods are highly structured, are developed from the Critical Incidents Technique (Flanagan, 1954; Bownas & Bernardin, 1988) job analysis method, and have demonstrated surprisingly high validities for structured interviews (Janz, Hellervik & Gilmore, 1986; Robertson, Gratton & Reut, 1990; Weekley & Gier, 1987).

The recent interest in the Situational Interview (SI) and the Patterned Behavior Description Interview (PBDI) methods is primarily because they are both behavior-based and have shown evidence of validity and reliability. Latham (1989) and Janz (1989) discuss further advantages and disadvantages of these methods. For instance, the SI is job-related, and it may be useful for positions which do not require past experiences due to a focus on intended behaviors. On the other hand, SI questions are stated hypothetically which may lead candidates to provide socially acceptable responses, and the scoring guide may be easily determined.

The PBDI looks at long-standing behavior as an indicator of future behavior, which is practical for positions which require past experience. However, this focus on past experiences may disadvantage some minority groups or women because they have not had the same opportunities as others. Another disadvantage with the PBDI is that it is difficult to develop and there are no procedures for developing scoring guides.

This paper describes the use of the Situational Interview and the Patterned Behavior Description Interview in a large, West Coast urban school district and how these methods were perceived by both interviewers and interviewees.

### The Critical Incidents Technique (CIT) Job Analysis Method

The CIT involves collecting behavioral examples of performance for a particular job. Supervisors and/or incumbents are asked to describe incidents which they have observed or performed on the job, including the actions that led to the incidents, the behaviors that were demonstrated and the consequences of these behaviors. Respondents are instructed to avoid using judgmental terms and to relay only the behaviors exhibited. These incidents are then grouped into dimensions. Similar situations within the dimension are compared for relative ranking in terms of the

desirability of the response to the situation, creating a foundation for evaluating candidates based on displayed behaviors in critical areas. The SI and the PBDI are then constructed from these data.

### The Situational Interview (SI)

In the SI, candidates are asked how they would respond to seemingly hypothetical incidents which could actually occur on the job. In fact, these situations are derived from the critical incidents collected for the job analysis. The SI is based on the premise that responses given will indicate the actual intentions of the candidates and that these are precursors of their actual behavior under similar circumstances (Latham, 1989; Latham, et al., 1980).

The SI was initially used in this school district to evaluate Custodian candidates in 1987 and was used again in 1989 with minor revisions.

Custodians are functionally supervised by Custodial Supervisors. Line supervision is provided by the principal of the school to which the Custodian is assigned. In 1987, critical incidents were collected from the Custodial Supervisors since they were most familiar with the actual job duties and requirements of the position. These incidents were grouped into five dimensions and SI questions were derived based on common incidents within each dimension.

In a series of pilot studies, SI questions were asked of both incumbents and candidates and their responses were recorded. These responses were then weighted (5) good, (3) mediocre, or (1) poor by their functional supervisors in a consensus meeting. A total of 31 questions were usable and two parallel forms of the test were developed containing 20 questions each, with some overlap of questions. Each form consisted of a rater booklet which included the 20 questions and a common example of a (5), (3), and (1) response; an answer booklet which contained additional responses and their weights; and a candidate booklet which contained only the 20 questions provided so that candidates could read along with the raters. An answer sheet was also developed on which the raters could circle the score of the response given by the candidate or write the candidate's response if it did not correspond with any of the answers in the answer or rater booklets. When the raters were unsure of how to score a response, after checking the answer booklet independently, they would discuss it with all the other raters in a group meeting following the interviews and a score was given on a consensus basis.

All raters were trained in the interview procedures and in the use of all three booklets, and viewed a mock interview videotape. They evaluated the candidates on the videotape and their scores were reviewed and discussed to assure that all raters assigned scores uniformly. One thousand eight hundred and eighty candidates were actually tested in 1987 using these procedures. The 1989 administration was similar except that some of the original questions were replaced with questions developed from critical incidents collected from principals/line supervisors who felt that they had not been given the opportunity to provide input for the original test. Eight hundred eight candidates were tested.

This large, West Coast school district has also conducted several other studies of the SI and has presented the results of these at various conferences over the past several years. These include a study of rater and ratee race effects (Lin & Manligas, 1988), the use of the Angoff method (Angoff, 1971) for pass point setting (Wieder & Lin, 1988), and a comparison of the SI with self-assessment techniques (Manligas & Lin, 1988).

#### The Patterned Behavior Description Interview (PBDI)

Like the SI, interview questions in the PBDI are derived from critical incidents as candidates are presented with work situations and asked how they have responded under similar circumstances. However, the PBDI is an interview method which is based on the assumption that past behavior predicts future behavior (Janz, 1989). Questions are followed by several probes designed to gather more in-depth information. Unlike most other types of interview formats which generally assess maximum performance (i.e. the type of behavior which would be exhibited under optimal conditions), the PBDI is designed to assess typical performance (Janz, 1989; Janz, et al., 1986). Thus, interviewees are asked to respond with the most recent, most difficult, most challenging, and/or least successful behavior examples.

#### The Current Study

The purpose of this study was to explore the applicability of both the SI and the PBDI in the selection of employees for entry and managerial level jobs. Specifically, would both the SI and the PBDI be perceived as better interviewing methods and more fair than conventional structured interviews? Would there be a difference between the SI and the PBDI in the eyes of the interviewers and the interviewees? It was hypothesized that both interviewees and interviewers would accept these two highly structure interview formats over the traditional structured interview.

#### METHOD

##### Situational Interview

##### Procedure.

The SI was used to test first-line cafeteria supervisors, an entry-level cafeteria supervisor position. Seventy-five non-repetitive critical incidents were collected from job incumbents. These incidents were grouped into common situations and then into dimensions. Situational questions were constructed from common incidents and tentative anchor responses were developed. The SI was then administered to incumbents in order to collect additional responses. These tentative responses collected from the job incumbents were then shown to the supervisors of the positions (acting as subject matter experts) and were weighted as: (5) good, (3) mediocre, and (1) poor. These responses were thus used as the scoring guide and candidates' responses were scored accordingly. The SI was the final part of a multiple hurdles examination.

### Subjects.

Thirty cafeteria employees who had taken and passed a written knowledge test and who met the minimum entrance qualifications took the SI. There were 28 females and two males; seven were White, six were Black, eight were Hispanic, and nine were other or unknown. All candidates had experience working in school cafeterias or other types of school food service facilities, such as nutrition centers or food packaging plants.

The seven interviewers participating in the project were management-level food services employees familiar with the positions, former incumbents, or personnel specialists who were thoroughly familiar with interviewing. Each candidate was evaluated by a panel of two raters, at least one of whom had an extensive food services background.

### Patterned Behavior Description Interview

#### Procedure.

The PBDI was used to test second-line cafeteria supervisors and data processing managers. For both examinations, incidents were collected from supervisors of the positions in the classes being tested. Seventy-five non-repetitive incidents were collected for the second-line cafeteria supervisor and 34 non-repetitive incidents were collected for data processing manager. These incidents were also grouped and combined into dimensions. Questions were derived from common incidents but were stated in such a way that candidates had to respond with examples of the most recent, most difficult, most challenging, and/or least successful resolution of a similar situation that had occurred to them. Initial questions were followed by probing questions designed to get an in-depth description of the event. Probing examples might include: How frequently does something like this occur? What events led up to the situation? How did you resolve the problem?

Candidates were rated using a behavioral rating scale authored by the critical incidents collected from incumbents. For both of these classifications, the PBDI was the final part of a multiple hurdles exam.

### Subjects.

For the second-line cafeteria supervisor classification, candidates were 25 cafeteria employees who had taken and passed a written knowledge test and met the minimum entrance qualifications. Twenty-three of these candidates were female, one was male, and one was unknown. There were six Black candidates, six Whites, and two Hispanics.

These candidates were evaluated by a total of six interviewers who were management-level food services employees familiar with the positions, former incumbents, or personnel specialists. Each candidate was evaluated by two raters, at least one of which had a food services background.

For the data processing manager exam, there were 11 candidates, all of whom met the minimum entrance qualifications requiring extensive



management experience and who had been successful on a Training and Experience Evaluation. Only five of these candidates responded to the questionnaire. All of these were male; four were white.

There were six interviewers in total for this examination, and three who responded to the questionnaire. All were obtained from outside the school district and were employed in positions similar to that for which we were testing, or were experts in the data processing field and were thoroughly familiar with this type of job.

#### SI and PBDI Interview Questions

Below, you will find an example of a critical incident, a SI question derived from that critical incident, and a PBDI question derived from it.

##### CRITICAL INCIDENT

This employee was devoted to his family. He had only been married for 18 months. He used whatever excuse he could to stay home. One day the fellow's baby got a cold. His wife had a hangnail on her toe. He didn't come to work. He didn't even phone in.

##### SITUATIONAL INTERVIEW QUESTION

Your spouse and children are sick in bed with a cold. There are no relatives or friends available to look in on them. Your shift starts in 3 hours. What would you do in this situation?

Possible responses (weighted)

- (5) I'd come to work since they only have colds
- (3) phone my supervisor and explain why I cannot come to work
- (1) stay home, my spouse and family come first

##### PATTERNED BEHAVIOR DESCRIPTION INTERVIEW QUESTION

Describe the most recent time your spouse and children were home ill and you were scheduled to be at work in a few hours.

- \* What were the circumstances
- \* Did you report to work? Why or why not?
- \* If you stayed home, how did you notify your supervisor?
- \* If you reported to work, describe a time when you did stay home to care for your family. When was this?
- \* How frequently do you stay home from work in a six month period?
- \* Does your staying home from work create an impact on the workload of others?

#### Feedback Questionnaire

Questionnaires were developed for all the SI and PBDI interviewees and interviewers. These consisted of questions which the respondents were asked to evaluate on a seven-point Likert-type scale. The questions focused on the respondents' impressions of the SI or PBDI method in a variety of different areas versus traditional structured interviews. A total of four questionnaires were developed, one each for interviewers and interviewees for the SI and one each for the interviewers and interviewees of the PBDI. See Figure 1 for the different questions used.

Insert Figure 1 about here

200

## RESULTS

Means and standard deviations were calculated for total responses to the particular questionnaire administered to each of the six groups; that is, raters for first-line cafeteria supervisor, ratees for first-line cafeteria supervisor, raters for second-line cafeteria supervisor, ratees for second-line cafeteria supervisor, raters for data processing manager, and ratees for data processing manager. See Figure 2 for the results of these analyses.

Insert Figure 2 about here

Responses to the PBDI questionnaires were then combined for comparison with the SI. Means and standard deviations were calculated for all respondents to the PBDI questionnaires, such that raters for second-line cafeteria supervisor and data processing manager were combined ( $N = 9$ , Mean = 4.78, SD = 1.48), and ratees for second-line cafeteria supervisor and data processing manager were combined ( $N = 31$ , Mean = 4.87, SD = 1.61). Means and standard deviations for responses to individual questions are also included in Figure 2. Any mean greater than 4.0 indicates a positive opinion of the PBDI or SI method over conventional interview formats.

All interviewers and interviewees responded to all questions with a mean greater than 4.0 except second-line cafeteria supervisor raters in response to two of the questions (see Figure 2). These are: is it easier or more difficult to ask questions using the PBDI method (Mean = 4.00), and did you feel more or less relaxed using the PBDI vs. traditional interview questions (Mean = 3.83).

T-tests were conducted to determine whether or not there were any significant differences between the means of 1) raters of the second-line cafeteria supervisors vs. raters of the data processing managers (entry vs. managerial level positions, using the PBDI), 2) ratees of the second-line cafeteria supervisors vs. ratees of the data processing managers (entry vs. managerial level positions using the PBDI), 3) raters of the first-line cafeteria supervisors vs. raters of the second-line cafeteria supervisors and data processing managers combined (SI method vs. PBDI method), and 4) ratees of the first-line cafeteria supervisors vs. second-line cafeteria supervisors and data processing managers combined (SI method vs. PBDI method). No significant differences were found in any of these analyses.

## DISCUSSION

The results indicate that all groups investigated in this study, overall, favored both the SI and the PBDI methods over conventional structured interview formats, supporting our hypothesis. That is, both interviewees and interviewers prefer these two highly structured interview formats to the traditional structured interview

However, interviewers for second-line cafeteria supervisor felt that there was no real difference between asking questions using the PBDI and traditional methods. Moreover, they felt less relaxed using the PBDI



over traditional methods. The reason for this might be that interviewers were required to ask many more questions than usual with the hope that two parallel forms of the interview might be developed for this particular exam and, therefore, may have felt that the whole interviewing process took too long.

The results also showed no statistically significant differences between the raters' and ratees' perceptions of the SI vs. the PBDI. This finding is consistent even between higher and lower level jobs. It appears that the SI and PBDI are viewed favorably by interviewers, and by candidates evaluated with these behavior-based methods. However, sample sizes for some of these groups were quite small and further studies will need to be conducted to replicate these findings.

Respondents also provided positive and negative comments regarding both interview methods. Positive comments about the SI included remarks that all candidates were treated uniformly, comparisons among candidates were easier to make, and that it gave the candidates a better chance to show off job skills. Some negative comments included that not all situations and responses were appropriate for the class tested, and that candidates may feel that they were not given an adequate opportunity to elaborate when responding to the questions.

Positive comments about the PBDI included the feeling that the PBDI was more thorough, detailed and specific, and that all candidates were asked the same questions. On the negative side, some respondents thought the questions were too long, very time consuming, and candidates were disadvantaged if they had not experienced the same situations about which they were questioned.

Overall advantages of the SI seem to be that it is easily scored, fairly easily developed, and very job-related. Disadvantages include its focus on the future and the use of hypothetical situations which lead to a concern that candidates will provide socially acceptable responses rather than responses which would indicate what they would actually do.

Overall advantages of the PBDI are that it focuses on past behavior and asks very in-depth questions, and is very job-related. Disadvantages include that it is not as easily scored as the SI, and there is the possibility that candidates will not have had the experiences in question. Latham (1989) and Janz (1989) have further discussions regarding the advantages and disadvantages of these methods. Latham (1989) also provides a chart comparing these two methods on 16 characteristics including job analysis method, evidence of utility, evidence of validity and reliability, scoring guide use, how questions are developed, and emphasis on behavior.

There are various factors which may confound the results of this study. None of the different classifications involved the same set of critical incidents, the interviews were developed by different analysts, there was no control for the number of questions asked of candidates in the different groups, and there were slight differences in the ways in which the raters were trained in the use of these methods. In order to obtain a completely unconfounded comparison of these methods, the ideal study

would involve evaluating all candidates using both the SI and the PBDI developed from the same set of critical incidents. Such a study would be very difficult to do, if not impossible, in a real-world setting. There would also be the need to test higher and lower level positions to further compare and contrast these methods.

In conclusion, a major difference between the SI and the PBDI is that they are based on different time frames in the candidates' experience. The SI is based on what candidates feel they will do in the future, and the PBDI is based on past experiences and behaviors. Consequently, it seems reasonable to assume that the SI method would be better suited to testing for entry or lower level positions because candidates for these positions will have fewer or no examples of pertinent past behaviors or experiences to relate. However, Robertson, et al., (1990) give an example of the SI used for administrative jobs.

On the other hand, because the PBDI is based solely on past behavior and experiences, it might be better suited for higher level or managerial positions for which candidates are expected to have substantial prior experience. These candidates would be able to describe how they have handled like situations in the past and may also be able to analyze the consequences of these experiences and how they might proceed in the future under similar circumstances. The PBDI method also seems to provide more detailed accounts of past experience, which would provide raters with more information prior to evaluating candidates for these higher level positions.

This large school district is currently developing an interview format that employs one SI question, one PBDI question, and one general interview question for each dimension being measured in the interview other than job preparation. This is similar to what is suggested by M. Campion, et al. (1988). There is a possibility that scoring comparisons could be made between these three different methods, although it will be difficult to assure that raters only focus on responses from one type of question when making this comparison. This approach may utilize advantages of both methods and diminish the disadvantages.

In sum, both the SI and the PBDI methods promise significant utility due to their differing temporal focuses on candidate behavior and their common high degree of standardization.

#### REFERENCES

- Angoff, W. H. (1971). Scales, norms and equivalent scores. In Thorndike (Ed.), Educational Measurement (pp. 514-515). Washington, D.C.: American Council on Education.
- Arvey, R. D. & Campion, J. E. (1982). The employment interview: A summary and review of recent research. Personnel Psychology, 35, 281-322.
- Bownas, D. A. & Bernardin, H. J. (1988). Critical incident technique. In S. Gael (Ed.), The Job Analysis Handbook for Business, Industry, and Government (pp. 1120-1137). New York: John Wiley & Sons, Inc.

Campion, M. A.; Pursell, E. D.; & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. Personnel Psychology, 41, 25-42.

Cronshaw, S. F. & Wiesner, W. H. (1989) The validity of the employment interview: Models for research and practice. In R. W. Eder & G. R. Ferris (Eds.), The Employment Interview: Theory, Research, and Practice (pp. 269-281), Newbury Park, CA: Sage Publications, Inc.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. Personnel Psychology, 42, 691-726.

Janz, J. T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. Journal of Applied Psychology, 67, 577-580.

Janz, T.; Hellervik, L.; & Gilmore, D. C. (1986). Behavior description interviewing: New, accurate, cost effective. Washington, D. C.: American Psychological Association.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), The Employment Interview: Theory, Research, and Practice (pp. 158-168). Newbury Park, CA: Sage Publications, Inc.

Latham, G. P. & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. Journal of Applied Psychology, 69, 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. Journal of Applied Psychology, 65, 422-427.

Latham, G. P. & Finnegan, B. J. (1987). The practicality of the situational interview. Unpublished manuscript. University of Washington, School of Business, Seattle.

Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), The Employment Interview: Theory, Research, and Practice (pp. 169-182). Newbury Park, CA: Sage Publications, Inc.

Lin, T. R. & Manligas, C. L. (1988, August). A comparative analysis of rater and ratee race effects in the employment selection interviews: Structured interview versus situational interview. Paper presented to the Academy of Management annual conference, Anaheim, CA.

Manligas, C. L. & Lin, T. R. (1988, June). The situational interview versus self-assessment: What can be done if candidates inflate their scores? Paper presented at the 12th annual conference of the International Personnel Management Association Assessment Council, Las Vegas, NV.

Robertson, I. T., Gratton, L., & Rout, U. (1990). The validity of situational interviews for administrative jobs. Journal of Organizational Behavior, 11, 69-76.

Weekley, J. A. & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. Journal of Applied Psychology, 72, 484-487.

Wieder, L. & Lin, T. R. (1988, June). Application of Angoff method in passing point setting for a situational interview. Paper presented at the 12th annual conference of the International Personnel Management Association Assessment Council, Las Vegas, NV.

Wiesner, W. H. & Cronshaw, S. F. (1988). The moderating impact of interview format and degree of structure on interview validity. The Journal of Occupational Psychology, 61, 275-290.

Figure 1

SI/PBDI Questions for Interviewers and Interviewees (consolidated\*)

- Q1. Is it easier or more difficult to ask/answer questions using the SI/PBDI method vs. traditional interview method?  
 7. . . . .6. . . . .5. . . . .4. . . . .3. . . . .2. . . . .1  
 much somewhat same somewhat much  
 easier easier more difficult difficult
- Q2. How thorough are the SI/PBDI questions vs. traditional interview questions?  
 7. . . . .6. . . . .5. . . . .4. . . . .3. . . . .2. . . . .1  
 much more same somewhat much  
 more thorough less less  
 thorough thorough thorough
- Q3. How thorough are the candidates'/were your responses to SI/PBDI questions vs. traditional interview questions?  
 7. . . . .6. . . . .5. . . . .4. . . . .3. . . . .2. . . . .1  
 much more same somewhat much  
 more thorough less less  
 thorough thorough thorough
- Q4. Did you feel more or less relaxed using the SI/PBDI vs. traditional interview questions? (Interviewers only)  
 7. . . . .6. . . . .5. . . . .4. . . . .3. . . . .2. . . . .1  
 much more same less much  
 more relaxed relaxed less  
 relaxed relaxed relaxed
- Q5. What was the quality (honesty, accuracy, relevancy) of responses/of your responses to SI/PBDI questions vs. traditional method?  
 7. . . . .6. . . . .5. . . . .4. . . . .3. . . . .2. . . . .1  
 much better same lower much  
 better quality quality lower  
 quality quality quality
- Q6. What is your overall feeling about the SI/PBDI vs. traditional method?  
 7. . . . .6. . . . .5. . . . .4. . . . .3. . . . .2. . . . .1  
 very positive neutral negative very  
 positive positive negative negative

\* 4 actual questionnaires were developed

Figure 2

## Questionnaire Results

	Q1	Q2	Q3	Q4	Q5	Q6	Total
First-line cafeteria supervisor (INTERVIEWERS)							
N	7	7	7	7	7	7	7
Mean	5.57	4.57	4.14	5.43	4.71	4.93	5.57
SD	0.90	1.05	1.64	1.18	1.28	1.37	0.98
First-line cafeteria supervisor (INTERVIEWEES)							
N	30	29	30		29	30	30
Mean	4.12	5.17	4.68		5.29	5.53	4.12
SD	1.47	1.27	1.23		1.09	1.02	1.50
Second-line cafeteria supervisor (INTERVIEWERS)							
N	6	6	6	6	6	6	6
Mean	4.00	4.67	4.50	3.83	4.33	3.83	4.00
SD	1.00	1.70	1.38	0.37	1.90	1.67	1.10
Second-line cafeteria supervisor (INTERVIEWEES)							
N	25	25	25		25	25	25
Mean	4.74	5.62	4.90		5.34	5.52	4.74
SD	1.64	1.32	1.43		1.32	1.44	1.68
Data processing manager (INTERVIEWERS)							
N	3	3	3	3	3	3	3
Mean	6.33	6.67	5.67	5.33	5.67	6.33	6.33
SD	0.47	0.47	0.47	0.94	1.25	0.47	0.53
Data processing manager (INTERVIEWEES)							
N	5	5	5		5	5	5
Mean	5.50	5.10	5.50		5.00	5.10	5.50
SD	1.00	1.28	1.34		1.26	1.43	1.12



ELECTRONIC DOCUMENT STORAGE

A CASE STUDY: JOB APPLICATIONS

IPMAAC PRESENTATION

JUNE 26, 1990

TED DARANY

JENNIFER FRENCH

## INTRODUCTION

For many years personnel selection practitioners have been in the process of automating a variety of activities. We have had notable successes in some areas, particularly in "applicant tracking". However, we have had a notable lack of productivity improvement in employing automation to manage our job applications. We have noted attempts to use technologies, such as Optical Mark Recognition (OMR) and Optical Character Recognition (OCR), with little success relative to job applications and resumes.

This paper focuses on a completely different approach: Electronic Document Storage (EDS). EDS does not translate a document into a text or coded form, but leaves it whole, storing it digitally. So when traditional data base techniques are linked with EDS, we may store and retrieve an electronic representation of the original document.

While EDS may offer significant advantages to several areas of personnel, we will direct our attention here principally to selection and use. as an example, a traditional job application.

## WHY EDS?

Many of us have had considerable success with automation. Most of us have automated major parts of our selection processes with an applicant tracking system. There is little question of the benefits of automated scoring, analysis, and candidate notification. Many of us have also employed automation for traditional word processing tasks and other areas such as budget analysis and selection research.

The reasons to pursue an electronic document storage system are the most traditional ones associated with data processing projects:

- to gain greater control over one's processes
- to be able to perform these processes in a more intelligent or flexible manner
- to save staff time and other costs.

EDS may accomplish all three of these objectives, for it can provide a means by which we can far more effectively control an original document, which is crucial to our process. Our goal in incorporating EDS for job application storage was most principally directed towards achieving greater control of the actual job applications; that is, we wanted to be able to be certain that we could retrieve specific job applications whenever we needed to.

A system that allows us to electronically store a document (in this case, a job application) allows us much greater flexibility as to the method by which we can retrieve that document. Most of us store paper documents in one particular way. Very few of us have cross references to such documents. But if a computer is responsible for storing and retrieving documents, we can include with the document any information we wish about the nature of

the document; that is, a document-specific data base. For example, we might manually store job applications in a folder with a label on it for a particular recruitment. Within the folder, the applications might be arranged alphabetically, by last name, or perhaps by social security number, or no particular order at all. With an electronic reference or index we can retrieve applications by many factors, such as name or social security number or city or state or ethnicity. We can even readily retrieve all the applications a given person has filed for multiple jobs. That kind of retrieval flexibility is virtually impossible (or extremely time-consuming) for a manual system, but relatively simple with an electronic system.

Of course, a third good reason to automate anything is to save taxpayers' dollars. Storing documents electronically has the potential for substantial savings in staff time involved with retrieval of the documents. Further, the electronic system may eliminate the expense of photocopying, for in many cases the document can be viewed on an appropriate display station instead of being printed.

#### WHAT IS EDS?

Electronic Document Storage (EDS) is a computer-based means of storing an image or picture of a document. It is not a conversion of a document into text or any other "intelligent" format. It's most directly analogous to an electronic photograph of the document. So EDS is a technology quite distinct from Optical Mark Recognition (OMR) and Optical Character Recognition (OCR). Many of us have thought to use either OMR or OCR to assist us with handling job applications. We will get to those possibilities in a moment. But first, let's be more specific about what OCR and OMR are since it may be important to keep the technologies distinct.

Many of us make frequent and very successful use of OMR for the scanning of test answer sheets. Our test scanning machines read the marks "bubbled" on the form by the examinee. The scanner senses the absence or presence of a mark and transfers the information into a computer, which decodes the mark based on its location on the form. This works wonderfully efficiently for forms such as written test answer sheets, oral examination and T & E evaluation forms, as well as job analysis questionnaires and surveys. OCR, on the other hand, is a technology which attempts to take typewritten or printed documents and turn the actual characters into their computer representation. OCR depends upon extremely sophisticated programming to convert a shape into a specific letter, number or symbol.

There have been a number of attempts to employ OMR and OCR to assist us in managing our job applications. That is, various employers have tried to scan job applications or resumes, either in the OMR or OCR manner. At this time, we feel that neither approach can be recommended for job applications. Completing an OMR job application is an extremely cumbersome task for the applicant. It takes many more pages and it's far more tedious a task and leads to far more frustration and errors than we can be tolerated. OCR breaks down since it requires a high degree of precision of the typed material and even then its ability to convert from what it scans is less than perfect. Of course, it's totally useless at reading most job applications

which aren't typed. We observe that it's at least theoretically possible for a complex program (usually referred to as artificial intelligence in this context) and specialized equipment to accurately scan and convert printed documents to computer-based text. Unfortunately, that observation has been makeable for 20 years. Some progress has been made but the degree of perfection currently existing leads to many critical errors, even with typewritten materials. EDS makes no attempt to convert the document from its physical representation into computer text. It simply takes an electronic photograph of the document and stores it efficiently. This doesn't help us with our desire to avoid having to key-enter the data from a job application for our applicant tracking data bases. We still have to do that. However, it does provide significant document storage and retrieval advantages.

## CASE STUDY: USING EDS FOR JOB APPLICATIONS

### The Importance of EDS for San Bernardino County Job Applications

Over the past 10 years, the rate of applications received in the Employment Division has increased from approximately 15,000 per year to the current 45,000. During this same 10-year period, the staff in the Examination Services Section, which supports application filing and retrieval, has only increased by 10%. We have been able to keep up with this increase in workload by effective automation in other functions of the Employment Division. Until recently we have made no effective use of automation for document storage and retrieval. This has resulted in a substantial increase in the staff resources we must use to maintain an adequate filing system. It has also resulted in an increasing number of misfiled applications, which itself leads to a significant amount of wasted staff time, large backlogs and embarrassing situations. The Examination Services Section now spends over 50 hours a week storing and retrieving applications (including the time involved in correcting mis-files, but not including supervisory time involved with "assistance".) That number of hours will be decreased through EDS.

We determined our maximum pay-off priorities to be:

- I. To have every job application retrievable at all times, throughout the recruitment and hiring process
- II. To be able to retrieve a given application by a variety of search parameters
- III. To save a substantial amount of the staff time involved in storing and retrieving job applications.

Based on these priorities, we sought an automated system which would:

- (1) Accurately scan, store, retrieve, display and print applications submitted for County job employment

- (2) Have sufficient storage capacity to allow for up to 25,000 applications to be on line
- (3) Allow for retrieval, using parameters that are as similar to our existing Data General (DG) applicant tracking data base as possible
- (4) Allow for high-quality printing of the stored application from any terminal hooked up to the applicant tracking DG computer
- (5) Provide an environment which would seem as similar as possible to the highly efficient and customized applicant tracking system on our DG.

### Description of EDS

Today's typical electronic document storage system contains the following components:

- a very high performance PC-based CPU, with a full page or two-page, high-resolution video display, main memory of 4 MB or more, disk storage of 300 MB to as much as 10 GB.
- a high speed (at least 10, preferably 30, page per minute) scanner.
- a high quality laser printer.
- software that allows all the equipment to work together, provides for an image data base and perhaps integrates with other computers.

Optionally, the system might include:

- a FAX card and software to send images directly.
- a color monitor equivalent to the monochrome one cited above.
- a file back-up system (tape)

### Practical Considerations

The image system used by the Employment Division of San Bernardino County cost \$69,000. Approximately half of that cost is in storage: disk drives and tape back-up for the job applications. The other half is in the other pieces of equipment and imaging software. This system can be scaled up or down in cost and capability. The principal variable is the image storage system. For an environment that would not require that all applications be on line all the time, a single optical disk drive would be satisfactory. That could lower the system costs from \$69,000 to \$35,000. Note that an optical disk based system would not require magnetic tape backup for the system and the deletion of that item was included in the cost estimate of \$35,000. Such a system would have essentially unlimited capacity for storage but the operator would have to load the right optical disk cartridge at

the point of retrieval for viewing or printing. Such a system would probably be more ideal for an organization which wanted to store something between 1,000 and 10,000 job applications, and one which would not require multi-user access. At the other end, a system could be configured having the ability to store approximately 150,000 applications on line for a cost of approximately \$150,000. For this cost, the system would include much more disk storage, as well as an additional work station, which itself would include an additional scanner and printer. It should be noted that there are not necessarily discrete points; that the systems can be scaled up and down; they can have a mixture of both magnetic and optical disks, and they permit great flexibility in networking view stations and even complete work stations.

Another practical consideration might be the availability of equipment and software and where trends are taking us. All the equipment described is readily available, although just barely. Much of it is state of the art, and, therefore, still a bit prone to unpleasant surprises and perhaps even unreliability. On the other hand, much of the technology that is being used in our image system has been used for slightly different purposes for some time now. So equipment reliability is probably no greater an issue here than for other computer equipment. But the integration of the various pieces, including the software, is what's newest. Most of the companies providing these kinds of systems are fairly small and fairly young. The business for these companies is growing explosively as the demand for these kinds of products is widespread in both government and the private sector. This will cause a continuous and fairly rapid reduction in the costs for a particular level of storage, and a continuous increase in the performance of the systems. For the above reasons we must recommend caution in embracing an EDS system at this time. We are convinced that the time was right for our jurisdiction and our experience leads us further to believe that many organizations would likely benefit from the capabilities right now. However, when considering automation or systemization priorities, we believe that EDS should definitely come after the implementation of a full-scale and efficient applicant tracking system and effective word processing. Based on our success in those two areas, and the need to handle documents more efficiently, we believe that EDS may be one of the most important automation tools throughout the 1990's.

Ref: IPMA Public Personnel Management, Volume 13, No. 4, Winter 1984, Pg. 451; "Computer Applications to Personnel (Releasing the Genie - Harnessing the Dragon)" by Theodore S. Darany



AN INVESTIGATION INTO THE INTERCHANGEABILITY OF  
ESSAY AND MULTIPLE CHOICE TESTS AS MEASURES OF WRITING ABILITY

Anne Forinash Friend  
Donna L. Denning, Ph.D.

Paper presented at the International Personnel Management Association Assessment Council Annual Conference, San Diego, California, June 27, 1990.

## Purpose

The purpose of this study is to examine the extent to which scores on multiple choice tests and essays are interchangeable as measures of writing ability by studying test performance of a group of applicants for the class of Management Assistant in the City of Los Angeles. It is one of several studies of testing writing ability planned or in progress for classes or groups of classes requiring different levels and/or types of writing as part of the duties of the class. The subject is important because almost all positions in City professional classes require some writing, and, for most, writing letters or reports is a frequent task. Current testing, described in detail below, consists of a multiple choice test, an interview, and an essay which is not graded separately but is given to the interviewers before the interview.

Use of multiple choice tests has the following advantages: (1) they can be scored quickly and easily; (2) they are very reliable; (3) once developed, the cost of administration and scoring is lower than that of other types of tests. The main disadvantage of multiple choice tests is that they are indirect, testing knowledge of some factors of writing competence (e.g., grammar or preferable usage), but not the ability to use that knowledge (recognition versus generation and organization).

The outstanding advantage of using an essay to test writing ability is that it is essentially a job simulation, testing writing ability by requiring applicants to write. The disadvantages of using essays are: (1) they take much longer to administer and to score than multiple choice tests; (2) scoring essays is more expensive than scoring multiple choice tests; (3) it is difficult to select and word subjects or prompts to evoke the desired type of response without giving any of the writers an advantage; (4) it is difficult to select and word subjects or prompts so that raters will not be influenced by a writer's opinion or knowledge; (5) reliabilities may vary considerably, especially when different subjects and/or readers are used, and are usually lower than those for multiple choice tests.

Therefore, evaluation of the extent to which essay and multiple choice tests are interchangeable as measures of writing ability is of great interest for employment testing. The time needed for scoring a test may mean losing a potentially superior employee who cannot wait to start work. And cost is always important.

## Literature Review

Most of the studies of measurement of writing ability have been done in the context of education. Applying this work in employment settings has both advantages and limitations. The context of these studies for use in education is almost always measuring the degree of success of previous training in order to determine its deficiencies and/or the amount and type of additional classroom work needed. The emphasis is on identification and diagnosis of problems with poor and borderline writers.

In the employment setting, the important question is, "Can this person do the job?", and the primary functions of the test are screening (either in or out) and ranking applicants, rather than channelling students or changing the teaching offered. In this context, face validity may override considerations of economy and reliability. Very strong evidence for any type of test which does not look like the job is required to convince supervisors and managers of its value. For example, a tensiometer measurement of upper body strength may be a better test for trainees for a highly physical job than carrying a ladder because the job sample also tests whether applicants know the best way to pick up the ladder, a skill taught after hiring; even employers who understand the difference often prefer the job sample. Also, in all too many situations, an employer must consider the possibility of having to convince a judge that a test which is not face valid is the better predictor of performance than a job sample.

The problem of both the cost and low reliability for traditional methods of scoring essays has been addressed by development of holistic scoring. In this type of scoring, a rater reads a piece of writing and assigns a point grade based on the overall impression of the piece; the rater may use a scoring guide describing each point on the scale, examples of writing at each level, or both. Raters practice together before scoring and criterion papers for each level may be identified, i.e., those for which all raters' initial scores agree. In this way, raters are both trained and 'calibrated' so that they begin with common understanding of the criteria and practice in using them the same way, reducing the number of unacceptable differences in initial scores; these differences are resolved before final scores are assigned.

Holistic scoring can be done effectively with short writing samples (150-300) words, cutting down administration time; essays can be read quickly and a single score assigned, which is faster than scoring by such methods as assigning points for using compound sentences or for 'good' paragraph structure or counting errors (methods sometimes referred to as atomistic scoring).

Using trained and calibrated raters and sufficient numbers of raters and essays, reliability and validity comparable to that of multiple choice tests have been reported. In an article describing the development and variations of holistic scoring, Cooper (1977) gives interrater reliabilities up to .98 for holistic scoring by trained raters, although reliabilities above .80 result from scoring using either three or more readings, two or more essays, or both.

Primary Trait Scoring, (Lloyd-Jones, 1977) describes a variation of holistic scoring in which one or several "primary traits" are defined. The traits should be easily evaluated and related to the desired response or to the purpose of the piece of writing. These primary traits are evaluated separately and summed to get the final score. The Doyle (1988) study is a good example of primary trait scoring in which acceptable levels of difference were set for each trait and for the final scores.

The Measurement of Writing Ability, (Godshalk, Swineford, and Coffman, 1966 [ETS study]) is the first major study of the extent to which multiple choice tests and essays are interchangeable in measuring writing ability and the first to use holistic scoring. In this study, the criterion used for most comparisons was the total score for five essays, three written in 20 minutes and two longer, each read five times (25 readings). For comparisons of essay scores or of test combinations which included an essay, the criterion was the total score for the five readings of each of the other four essays (20 readings). The essays were graded on a three point scale, and two were also graded using a four point scale; raters were experienced in reading essays, but not in holistic scoring and did not practice together. All of the essays were read by 25 raters working in the same place during a five day period. Several times in the first two days, selected sample papers and all of the scores given to them were distributed for assistance in raters' self evaluation. No other training was given. Mean essay scores varied according to the topic of the essay; however, these variations did not affect the criterion because the total of all the scores was used.

The score reliability achieved for the 25 reading criterion was .84. This criterion was used for calculating correlations with various numbers of readings of the individual essays. Reliability for two readings of one of the short essays graded on the three point scale was .38. The correlations given here have been selected because they are data for common types of testing and scoring procedures. Correlations were also calculated for the criterion vs. scores for six short multiple choice tests of different types of items, with the SAT verbal score, and with scores on two interlinear (editing, not multiple choice) exercises. Correlations with the writing criterion for the multiple choice tests were: paragraph organization (putting sentences in the correct order), .46; English usage, .71; sentence correction, .71; paragraph completion (Which of the following sentences best fills the blank?), .57; error

recognition, .59; construction shifts (If a part of a given sentence is changed, e.g. coordinate conjunctions, choose the answer giving the alternative wording required by the change.), .65. For a test battery consisting of sentence correction, paragraph completion, and construction shifts, correlations of the composite score with the criterion ranged from .70 to .73 for four groups of writers; when a short essay, graded on either a three point or a four point scale, was substituted for the construction shift section, correlations increased slightly to a range of .73 to .75. The authors conclude that multiple choice tests are a valid test of writing ability, and that a combination of multiple choice items and an essay is an even better predictor of writing ability, but is unlikely to be enough better to justify the expense unless the effects of face validity are very important. In this study validity coefficients for an essay alone read fewer than three times were not comparable with those of multiple choice tests ( $r$ 's= .45-.48 for two readings).

Breland and Gaynor (1977), in a very similar study which used six calibrated raters, two readings per essay, a six point scale, and a scoring guide describing each point on the scale report score reliability of .51 for the essay. Three different versions of the Test of Standard Written English (TSWE) had correlation coefficients of .72, .74, and .69 with the composite essay score. The authors attribute the improvement in essay score reliability compared with the 1966 ETS study to several factors. Staff working with the College Board reported to the authors that in the time between the ETS and Breland-Gaynor studies (1) essay questions were gradually standardized; (2) test questions were pre-tested; (3) time allowed for responses is more optimal; (4) more is known about topics and responses to be expected; (5) more use is made of scoring guides based on a sample of the essays; (6) readers are more carefully instructed; (7) reading sessions are more carefully monitored. This study did not examine the effect of combining essay and multiple choice items in one test.

A subsequent study (Benton and Kiewra, 1986) used the same scoring guide and six point scale used in the Breland and Gaynor study and compared scores on two essays, TSWE, and four types of multiple choice items designed to test organizational skills at different conceptual levels (requiring test takers to order mixed up letters in words, words in sentences, and sentences in paragraphs, and to sort groups of nine sentences into three paragraphs). Stepwise multiple regression analysis using combined essay scores as the criterion, indicates that the TSWE explains 22% of the variance ( $r = .47$ ) and a composite of the four organizational skills scores explains an additional 7% of the variance ( $r = .54$ ) in writing ability.

The Doyle study mentioned previously is an excellent example of primary trait scoring, of the difference in approach needed for employment testing, and of combining research with operational employment testing. In educational testing, whether the tests



are used for entry level qualifications or for tracking or placement, information gained from the test is used for direct counselling or placement of the majority of test takers, even those who score in the lower ranges; therefore, although expense is important, it can be justified based on the fact that the test benefits both the institution and most of those taking it. The test described in the Doyle study presents a quite different situation, and one very common in employment testing. Hundreds of applicants were expected to compete for places on a list to be used to fill an estimated 15 positions as Public Information Specialist. Considerations entering into the final design of the examination were: (1) the job centers on writing ability; (2) a wide range of qualifying experience was accepted, creating a very diverse applicant group; (3) the cost of testing many applicants relative to the number of expected positions is very high; (4) applicants dissatisfied with their scores have the right to challenge either the test or the scoring through an appeal process.

The procedure decided on was to use a four part multiple choice test weighted 50% and an essay test weighted 50%. The essay test consisted of a 300 word news release and a 150 word public service announcement, both to be written based on information provided. In order to reduce costs, the essays were scored only for those passing the multiple choice test. In order to provide documentation in some detail in the event of challenges, both essays were scored on four factors using a four point scale (0-3) for each factor. Permissible differences between raters were: factor scores, one point; each essay, two points; combined essay scores, three points. Each rater could give each essay a maximum of 12 points.

Correlations between the multiple choice test and the essay test scores were low, partly because of restriction in range caused by scoring only essays written by those who passed the multiple choice test and partly because the essay factors took such global qualities as "selection of information to include" and "overall effectiveness" into account. The essay will be retained in some form in future examinations. Correlation of the "overall effectiveness" factor with the total the other three was .81 and with the total of the four factors was .89; some savings could be made through giving a single score for each essay, but probably not enough to offset the disadvantages of losing documentation.

In a study of its Educational Placement Test, the California State University System (Bianchini, et al., 1986 [CSU study]) found that the holistic essay scores had the following correlations with multiple choice tests: reading comprehension, .53; sentence construction, .56; logic and organization, .52. The correlation with a multiple choice test battery which included all three parts was .58. Intercorrelations between the multiple choice parts are: reading-sentence, .78; reading-logic, .83; sentence-logic, .76. The committee recommended retaining the essay, believing that the difference in ranges of the two



sets of correlations suggests that the essay measures different skills from those measured in the multiple choice test. However, they also note that a relatively high or low score on the multiple choice test alone is a good predictor of performance on the entire test.

For the present research, essays were scored with a modification of the holistic scoring method used in the California State University System English Placement Test. The version used was created by Charles Moore (California State University Sacramento) and adapted for employment testing by Richard Honey (Commission on Peace Officer Standards and Training, California State Personnel Board). Brief descriptors of each level of a six point scale were given to the four participating raters. A number of papers was read by the entire group of raters, and a criterion paper exemplifying each level of the scale was identified when all raters agreed on the initial score; copies of the criterion papers were then given to each rater and were used for reference while scoring the other essays.

Discussion of initial scores during identification of criterion papers is extremely valuable in training raters to focus on the writing itself, rather than the content. That distinction is essential when the intention is to evaluate writing ability alone, the purpose for which holistic scoring was developed; it may also be more difficult to maintain when reading essays on a general topic which is face valid for an employment test than for a topic evoking the writer's personal experience because anyone who is familiar with the job is likely to have a personal concept which includes content of "a good response" to a prompt, however general, related to the work.

Each essay was read by two of the four raters, all members of the City Personnel Department Research section at the level of Personnel Analyst or higher, (approximately 150 essays read by each rater). To balance idiosyncratic effects, each of the six possible pairs of raters read approximately fifty essays. Each rater scored each paper independently. The scores were compared by a third rater and, when scores for a paper were more than one point apart, the original raters rescored the papers independently. Disagreement after rescoring was resolved by discussion between the raters so that none of the final scores differ by more than one point for any applicant. The applicant's final score is the sum of the two raters' final scores.

## Method

### Class Selection

The class of Management Assistant was selected for one of the first studies. The minimum qualifications for applicants, as presented in the examination announcement, are:

1. A bachelor's degree from a recognized four-year college or university and achievement of a passing score on the qualifying Management Assistant abilities test and completion of an Advisory Writing Problem; or
2. A master's degree from a recognized college or university and completion of an Advisory Writing Problem.

This class is desirable for study for four reasons. (1) As the City's general entry-level professional class, some writing is one of the duties of almost all of the positions; (2) the requirement that applicants have either a bachelor's or a master's degree ensures some similarity in educational background and that the applicants have some experience in writing; (3) weekly testing and the tests used make it possible to collect data for a large group of applicants in a relatively short time; (4) since the examination is given on an "Open" basis, only, with no experience requirements, the applicant group is more diverse than those for most City examinations.

### Applicant Testing

Applicants are asked to bring their diplomas with them to the test. Each applicant scheduled for an interview is asked to take the diploma to the background section whenever there is time during the day; this section verifies that an applicant's degree is from an accredited school.

Applicants are tested using the following procedure:

1. All applicants complete the writing exercise. The prompt is a broad question related to office work and does not require specific knowledge. The resulting essays have a length of approximately 100-400 words.
2. Applicants with bachelor's degrees take the multiple choice test, a high level test of aptitude consisting of 100 items of four types: vocabulary, verbal analogies, basic arithmetic, and number series. It should be noted that the verbal items do not directly test knowledge of grammar or correct English usage. The qualifying score is 65.

3. Applicants with master's degrees and those with bachelor's degrees who pass the multiple choice test are interviewed. For each applicant interviewed, the applicant's essay is given to the interview board for review before the interview starts. Interview boards consist of two members selected from City employees at the level of Senior Management Analyst or higher. Interview board members change from week to week, and repeating interview board members usually serve with different partners.

Five factors, background, personal qualifications, interpersonal skills, written communication skill, and oral communication skill, are considered in arriving at the final score. Although they are not given specific individual weights, space is provided on the worksheet where interviewers can indicate ratings of each factor separately. After every interview, each of the interviewers independently assigns a score to the applicant using five point increments on a scale ranging from 65 to 95; then the scores are compared; if there is a difference of ten or more points, the scores are discussed; scores are usually adjusted, so that the final difference is not more than five points, but this is not an absolute requirement. An applicant's final score is the average of the two interviewers' scores, rounded to a whole number. The minimum passing score is 70, and all failing applicants are given a score of 65.

#### Sample

Over a period of six weeks, data were collected for all Management Assistant applicants, a total of 307. Since applicants who fail the written test do not go on to the interview and those with master's degrees do not take the multiple choice test, there are three groups of applicants: (1) those with multiple choice and essay scores, only (N = 128); (2) those with essay and interview scores, only (N = 67); (3) those with multiple choice, essay, and interview scores (N = 112). This results in considerable variation in the numbers of cases reported for different analyses. However, the number of cases for each analysis is approximately equal to the sum of at least two of these three groups and is large enough to permit reliable analysis.

## Instruments

Five measures, all attained in or adapted from the actual examination, were used in this research. They are:

1. the score on the complete multiple choice test;
2. a verbal multiple choice test score created by rescoring the multiple choice test on only the vocabulary and verbal analogy items (60 items);
3. a holistic essay score (scoring method described in the Literature Review above);
4. an "Interview Writing Skills" score consisting of the average of the ratings given by each of the two interviewers on the "written communication" factor in the interview;
5. the final interview score consisting of the average of the final ratings given by each of the interviewers.

For applicants who were interviewed, scores on the interview factors and tentative and final overall scores given by each interviewer were collected from the interview worksheets and transferred to a data sheet. These data sheets, score sheets from the multiple choice test and records of the raters' essay scores were then used for computer data entry.

## Data Analysis and Results

A summary of test results for the total sample is given in Table 1. The mean for the complete multiple choice test was 63.34 (range 16-97, 100 items); for the verbal multiple choice test, 39.57 (range 10-59 60 items); for the essay, 7.21 (range 2-12, 1-12 point scale); for the interview writing skills factor, 83.46 (range 65-95, full range used); and for the final interview score, 83.67 (range 65-95, full range used).

Results of t-tests comparing the means on each of the five research measures for candidates who passed vs. failed the complete multiple choice test are shown in Table 2. The significant ( $p < .0001$ ) differences in scores on the complete multiple choice test (pass, 77.00; fail, 50.78) and verbal multiple choice test (pass 48.24; fail, 31.60) simply reflect the fact that the criterion for dividing the applicants into groups was whether they passed or failed this test. However, the difference in mean scores on the essay (pass 7.73; fail 6.46;  $p < .0001$ ) for applicants grouped by their performance on the multiple choice test tends to support the hypothesis that there is some degree of similarity between essay and multiple choice scores.

## Reliability

Correlation of the individual interviewers' scores on the written communication factor for each applicant interviewed resulted in a coefficient of .73 ( $p < .0001$ ). The correlation for the tentative overall scores, assigned before discussion between the interviewers, is .88 ( $p < .0001$ ). Correlation of final interview scores is .95 ( $p < .0001$ ). Thus, in each instance the agreement between raters is substantial, with the overall scores demonstrating greater reliability than individual written communication factor scores.

Turning to the reliability of the holistic essay scores, the correlation for raters' scores before revision or discussion is .61 ( $p < .0001$ ). After adjustment using the Spearman-Brown formula for estimating the correlation of partial scores, the reliability for scores before revision or discussion is .77 ( $p < .0001$ ). The correlation of raters' final scores (after discussion) is .78.

## Relationship between Tests

Correlations were computed comparing performance on interview and written test scores (Table 3). Correlations between the various interview factors and the final interview score are also given.

Correlation coefficients for the both the complete and the verbal multiple choice tests with all of the interview factors are low and not significant. Correlations for the holistic essay score and four of the five interview factor scores are larger and statistically significant ( $r$ 's = .25-.28;  $p < .001$ ). Those for the essay score with the written communication factor score and the final interview score are considerably larger ( $r = .44$  and  $r = .36$ , respectively;  $p < .0001$ ).

Intercorrelations of scores on all tests used in the research are given in Table 4. The critical datum required for testing the hypothesis that multiple choice tests and essays can be used interchangeably to test writing ability is the correlation between the verbal multiple choice subtest score and the holistic essay score. This correlation,  $r = .43$  ( $p < .0001$ ), is moderately high, suggesting substantial, though by no means complete, overlap of the two types of test. Furthermore, the correlations for the complete multiple choice test, the verbal multiple choice test, and the interview writing skills scores with the holistic essay score are very consistent (.43-.46;  $p < .0001$ ).

The fact that the correlation between the essay score and the complete multiple choice test score is slightly higher ( $r = .46$ ) than the correlation with the verbal multiple choice test score is partly explained by the greater length of the complete multiple choice test. However, both the verbal and the numerical items on the test include relational items which test logic and organizational ability.



## Comparison with Prior Research

The correlation of the verbal multiple choice test score with essay score derived from the present research ( $r = .43$ ) and the correlation for the logic section of the multiple choice test and essay ( $r = .52$ ) used in the CSU study were converted to Fisher's  $z$ 's and the difference tested for significance; the difference is not significant. This particular comparison is made because the verbal multiple choice items used in the present research includes verbal analogy items and has no items similar to those used in the other CSU study multiple choice test (Reading Comprehension [ $r = .53$ ] and Sentence Construction [ $r = .56$ ] sections). Results of the CSU study also indicate that although the essay adds to the reliability of the multiple choice test, it is critical only in the range near the pass point; the multiple choice test alone identifies students who perform in the good to very good range or who clearly fail.

## Conclusions and Recommendations

There is a definite relationship between holistic essay scores and multiple choice tests of verbal ability, even when the multiple choice test is a general verbal test which does not specifically test knowledge of English grammar or usage. The correlation ( $r = .43$ ,  $p < .0001$ ), although quite high, is not sufficiently high to support using this multiple choice test as a measure of writing ability.

At the same time, the similarity of the results of this study to the CSU study logic subtest-essay correlation suggests that, following that precedent, using a multiple choice test could give results usable for screening out applicants who do not write well enough to write acceptable letters, memos, reports, or similar material. A frequency distribution using a cut score of 7 points (maximum, 12) for the essay and 50% for the verbal multiple choice test shows that 41% passed both tests, 20% passed the essay but failed the multiple choice test, 19% passed the multiple choice test but failed the essay, and 21% failed both tests. To ensure that no one who wrote an adequate essay would be screened out with the multiple choice test required lowering the cut point to almost 25%, the chance score for guessing alone.

Although the results of this study show a relationship between essay and multiple choice test scores, the correlation is substantially lower than those reported for multiple choice tests designed to test either knowledge of correct English usage or the ability to identify the correct or preferable word(s), or sentence from several alternative. The ETS study reports validities ranging from .57 to .71 for these types of items (a lower correlation, .46, is reported for paragraph organization, which is a more relational type of item). Breland and Gaynor report correlations of .72, .74, and .69 for administrations of three different forms of the TSWE vs. a composite essay score, and Sinclair (1989) reports correlations ranging from .59 to .64



for the TSWE vs. six different essay scores. Research to test the hypothesis that multiple choice tests designed to test knowledge of or ability to use standard English are more likely to be interchangeable with essay tests as a measure of writing ability is now in progress.

In addition, our experience is that, after training, it was possible for raters to read essays of this length rapidly (three of the raters estimate 5 minutes per essay scored, average) and achieve initial scores no more than one point apart more than 80% of the time. (Fewer than 10 [0.7%] essays received scores three points apart, and none had a greater difference.) It is possible that giving members of interview boards a brief orientation to holistic scoring techniques and a list of short descriptors suited to the material they are to read would permit them to make more productive use of the very limited time they are given to read the essays before the interview starts. When an independent evaluation of writing ability is important, such procedures might also lessen the influence of the content of the essay, the halo effect of the interview, and trivial errors.

# APPENDIX

## TABLE 1

### TEST RESULTS, TOTAL SAMPLE

	N	Mean Raw Score	SD	Range	Scale
Complete Multiple Choice	238	63.34	16.04	16 - 97	1 - 100
Verbal Multiple Choice	238	39.57	10.91	10 - 59	1 - 60
Essay Scores	302	7.21	2.05	2 - 12	1 - 12
Interview Writing Skills	175	83.46	6.41	65 - 95	65 - 95
Final Interview	179	83.67	7.26	65 - 95	65 - 95

**TABLE 2**  
**TEST RESULTS BY PASS/FAIL COMPLETE MC TEST**

**PASS**

Test	N	Mean	SD	Range
Complete Multiple Choice	114	77.00****	8.50	65 - 97
Verbal Multiple Choice	114	48.24****	5.78	29 - 59
Essay Scores	179	7.73****	2.00	3 - 12
Interview Writing Skills	175	83.46	6.41	65 - 95
Final Interview	179	83.67	7.26	65 - 98

**FAIL**

Test	No.Cands.	Mean	SD	Range
Complete Multiple Choice	124	50.78****	9.86	16 - 64
Verbal Multiple Choice	124	31.60****	8.05	10 - 47
Essay Scores	123	6.46****	1.90	2 - 12
Interview Writing Skills	NA	NA	NA	NA
Final Interview	NA	NA	NA	NA

\*\*\*\* p<.0001

TABLE 3

## CORRELATIONS OF WRITTEN TEST AND INTERVIEW SCORES

Test	N	Back-ground	Pers. Quals.	Inter-pers.	Oral Comm.	Written Comm.	Final Interview Score
Complete Multiple Choice	112	.08	.05	.11	.14	.11	.13
Verbal Multiple Choice	112	-.16	-.16	-.03	-.08	-.10	-.15
Essay Scores	177	.25***	.28***	.27***	.26***	.44****	.36****
Final Interview Score	177	.82****	.89****	.82****	.87****	.73****	1.00

\*\*\* p&lt;.001

\*\*\*\* p&lt;.0001

TABLE 4

## INTERCORRELATIONS OF WRITTEN TEST SCORES

Test	Complete Multiple Choice	Verbal Multiple Choice	Essay Score	Interview Writing Skills
Complete Multiple Choice	1.00			
Verbal Multiple Choice	.90****	1.00		
Essay Score	.46****	.43****	1.00	
Interview Writing Skills	.11	-.11	.44****	1.00
Final Interview	.13****	-.15	.36****	.73****

\*\*\*\* p&lt;.0001

## BIBLIOGRAPHY

1. Bianchini, J., et al. (1986). Report of the English placement test evaluation committee. California State University, CSU Long Beach.
2. Breland, H. and Gaynor, J. (1977). A comparison of direct and indirect assessments of writing skill. Journal of educational measurement vol. 16, no. 2, Summer, 1979.
3. Cooper, C. R. (1977) Holistic evaluation of writing Evaluating Writing, ed. Cooper and Odell, State University of New York, Buffalo, 1977.
4. Doyle, E. (1989). The testing of writing ability: an experimental approach, the results and conclusions. New York State Department of Civil Service, Albany, NY.
5. Godshalk, E. F., Swineford, F., and Coffman, W. (1966). The measurement of writing ability. Research Monograph No. 6. New York, NY.. College Board Entrance Examination Board.
6. Lloyd-Jones, R. (1977) Primary trait scoring in Evaluating writing, ed. Cooper and Odell, State University of New York, Buffalo, 1977.
7. Noreen, R. (1989). California State University focus on English: English placement test. California State University, CSU Long Beach or CSU Northridge.
8. Sinclair, N. (1989). Assessing the writing of teacher candidates: Connecticut's method of holistic Assessment. Division of Research, evaluation and Assessment, Connecticut State Department of Education
9. Wangberg, E. G. and Reutten M. K. (1986). Whole language approaches for developing and evaluating basic writing ability. Lifelong Learning, Vol. 9, No. 8, June, 1986.
10. Wansor, C. E., (1986). Assessing writing ability: what are the issues, approaches? NASSP Bulletin, January, 1986.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000  
by: Kay Barrow

COLORADO DEPARTMENT OF PERSONNEL

USING TODAY'S TECHNIQUES AND TECHNOLOGIES  
TO PREPARE FOR ASSESSING WORKFORCE 2000

Why do we need to prepare for assessing Workforce 2000? Because the world is changing! New technology assaults us before we have mastered the old. Our work force is becoming increasingly diverse. We are experiencing a work force which includes older workers, more women, more minorities, and more immigrants. We have been instructed by labor statistics to expect a labor shortage in the near future and a greater competition for the qualified employees in the job market. There is a wave of competing demands of home and family which must be countered by flexibility in the work place. And, finally, the demand for increased productivity from a decreased work force will exacerbate.

These are the reasons we must prepare NOW for assessing Workforce 2000!

The State of Colorado has been in the process of critically reviewing its personnel system to determine how best to meet the demands which will be placed upon it when the impact of Workforce 2000 becomes a reality. The bold truth that confronts us is that the current personnel selection system is not adequate to meet the changes which are coming.

Currently, the Colorado Constitution requires that jobs in the state personnel system be filled based on merit and fitness as determined by competitive tests of competence. Under present practice, different exams are developed and given for each of approximately 1,600 job classes. Additional tests are often constructed for specific jobs within those classes as well.

An individual seeking employment in state government must complete a separate application form for each currently vacant job in which he or she is interested. Applications cannot be placed on file for consideration when future openings occur.

This system is widely perceived as unwieldy by both applicants and managers who have jobs to fill. Its major drawbacks are:

- It is reactive and frequently requires a "start from scratch" approach when an opening occurs. This in turn leads to a lengthy and cumbersome process for both applicants and managers seeking to fill jobs quickly.
- Because in most cases applications are accepted only during a short recruiting "window" when a vacancy exists, the quality and quantity of applications received may be restricted. Well-qualified potential candidates may not learn of a job opening in time.
- The current process requires an applicant to keep checking with various agencies to see whether openings have occurred. An applicant cannot submit one central application.



CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

- Because this reactive process produces pressure to fill vacancies quickly, affirmative action considerations may receive short shrift.
- This process is labor intensive. It is costly to design and administer hundreds of separate tests, and to separately screen hundreds of applications for each of them.

Because of the unwieldy, cumbersome and relatively inflexible selection system, early in 1989, Colorado developed an initiative called New Directions which will enable us to meet the needs of our changing and highly competitive workforce.

The challenge of New Directions in Selection is to develop an equitable, streamlined process that will:

- Provide a user friendly selection system that will allow applicants to easily apply and equitably compete for state jobs.
- Provide state managers the ability to make the best selection decisions in a timely manner from a highly qualified applicant pool.
- Provide an equitable process for promotion within the state personnel system.
- Provide a uniform automated selection process to be used by all state agencies.
- Reduce the cost and human resources necessary to administer the selection process statewide.

The benefits of New Directions are numerous. Some of the most evident are:

- An obvious acceleration of the overall employment process through having a ready supply of screened candidates at the time openings occur.
- Quantity and quality of applicants will increase through maintaining a continuous outreach and screening process. Currently, selections are made from applicant pools generated over a short time period.
- Applicant pools will be available for constant review by Selection Center personnel to determine if the desired numbers and types of skills are represented. This will enable a better concentration of recruitment efforts directed at specific need categories. It will also enable a better attention to affirmative action by intensifying recruitment in specifically defined areas where underrepresentation has been readily identified.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

- Department heads and their managers may review applicant files within their appropriate job-family groupings. This will provide further assurance that screening procedures and tools are, in fact, providing qualified applicants.
- Impacts on workloads and costs will also be realized. Individual job-opening announcements and advertisements may be eliminated except as needed to replenish the applicant bank. Selection Center and decentralized-agency staff time should be more effectively deployed instead of being involved in a repeated start-from-scratch recruiting effort in response to every job opening.
- Job applicants will receive broader consideration for across-the-board state employment opportunities rather than being evaluated only for the specific openings for which they apply. This will lead to a better public image of the state hiring process.

The new selection process will generally work as described below:

- A job seeker will complete one state application form, a job interest list, and an inventory of specific skills, education and experience.
- He or she will then take a general abilities test, if appropriate, and any other predetermined tests needed.
- Information about this applicant produced by these forms and tests will be stored in a computerized data base.
- Job profiles for different jobs in the state personnel system will also be stored in the automated data base.
- As vacancies occur, agencies will identify the kinds of jobs they are and whether any special knowledge, skills or abilities are required for the particular jobs they need to fill.
- If no special knowledge, skills or abilities are needed for a specific opening, the job profile and general abilities areas will be matched with the three top scores, skills inventories, and job interests of applicants on file, and the three best-matched applicants will be referred for consideration for the job.
- If special knowledge, skills or abilities are needed, the current applicants meeting the job profile will be given additional tests for the specific needs, and the three best-matched candidates will be referred for the job.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

In order to bring about the changes that are planned for selection, it was necessary to first make changes in the way that the Colorado Department of Personnel's Selection Center managed its exam development and administration activities. No additional staff were added for implementing change, so current staff spent considerable time planning ways to accomplish major changes while still conducting business as usual.

The solution to accomplishing all of the planned changes was to divide the proposed changes into manageable chunks and use a project approach to dealing with those chunks. The plan which emerged was a three-stage, five year plan, beginning with four projects that were top priority and ending with four other projects which were suitable for postponement to a later stage.

The four initial projects were: 1) Completion of new, well documented job analyses identifying critical knowledge, skills and abilities for all classes; 2) streamlining the current exam administration process and reducing analysts' workload, to free up time for the new activities; 3) ensuring that there was an enhanced automated applicant tracking system for the mainframe and that there was microcomputer capability for item banking; and 4) identification of test needs systemwide, so that a priority list for new exam development could be established.

Another project, development of an item banking system, emerged as so critical to the proposed changes that it could not be deferred until phase two. This fifth project got underway 6 months after the other four.

The three remaining projects, development of a New Directions process for promotional exams, formal cost-benefit analyses and marketing/education were deferred until phase two, although elements of each of these were discussed and dealt with on a limited scale. Phase three will be the full implementation of planned changes and documentation tasks which will include all necessary revisions to rules and procedures.

Team leaders and members for each of the projects were current staff "volunteers" with interests in the specific areas. A project coordinator was appointed to ensure that the project functions would avoid any duplication of effort. All personnel staff from the decentralized agencies were invited to participate in any of the projects too, so that agency personnelists could have the opportunity to participate in the implementation of changes that would have major impact on them in the future.

By using the project approach, with team leadership and coordination provided by department and agency staff, the Selection Center is accomplishing change without the disruption of day-to-day business.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

A brief description of each project follows for your information:

- PROJECT 1: JOB ANALYSES AND KNOWLEDGE, SKILLS AND ABILITIES (KSA) PROFILES

The New Directions selection methodology requires that applicants' test scores and other qualifications be matched against the KSA's required to do the job for which they are applying. This means that, for every job in the state, there must be a KSA profile against which the applicants' test scores can be compared.

The initial goal of this project is to develop a job analysis for about 90 state jobs in the professional services area. The profiles will be established within the job analyses. When this initial phase of the project is completed, we will use the results of the project to collect job analyses and profiles for the other approximately 1400 jobs in the state.

- PROJECT 2: STREAMLINING THE CURRENT EXAM ADMINISTRATION PROCESS

Over the years, there has been much discussion concerning the time it takes to fill positions in the state personnel system. Managers have complained that the process takes too long and they have positions vacant too long and may end up losing the position. There is also the concern that the better candidates get other jobs while waiting for a state job to be filled. This seems to be a systemwide problem but greater in some agencies than others.

Second, the current Director of the Department of Personnel believes that the Department should not be involved in the day-to-day provision of services to agencies (i.e., classification and testing). This Director would rather decentralize all of these functions to the agencies.

Finally, in order to accomplish many of the projects and tasks involved in New Directions, it was decided that those individuals involved in exam administration (exam analysts and technical support staff) needed to either reduce the amount of work involved in selection and/or look at the selection process differently.

To accomplish these three related objectives, the selection streamlining taskforce was initiated.

The implementation plan was designed to include three phases: Phase I focused on methods of streamlining and reducing the exam administration activity in the Selection Center of the Department of Personnel.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

Thirteen different steps were identified to accomplish this goal. These steps included such things as appropriating other eligible lists to fill positions, reducing the number of multi-step exams, eliminating agency-specific testing and streamlining general use class testing.

Phase II is the transition phase of taking on additional New Directions project activities as staff time is freed up from exam administration.

Phase III represents the ongoing activities following full implementation of New Directions. The primary emphasis is planned to be assistance and consultation to agencies and systemwide oversight of the selection function.

- PROJECT 3: AUTOMATION

One of the key and necessary components for the success of New Directions in Selection is a statewide automated applicant and examination tracking system. The Department of Personnel currently operates an automated applicant tracking system on the state's IBM Sierra 3090-400E mainframe located in the Department of Administration General Government Computing Center (GGCC).

The first step in meeting the stated objectives of the project was to form a task force made up of individuals from the private sector (IBM, Martin Marietta and U.S. West), other state agencies (Administration, Highways and Revenue) and the Department of Personnel. The mission of the task force was to investigate the types and kinds of software available for applicant tracking systems, evaluate the products and make a recommendation to the Director of Personnel.

A list of needs for an automated applicant tracking system was developed and used to evaluate the software packages available for mainframe, mini-computer, and PC systems.

A report evaluating the different software/hardware options was presented at the final task force meeting. Based on the review of the findings the task force recommended to the Director of Personnel that the current automated system be retained and enhanced to meet the additional needs resulting from the proposed improvement to the selection process as well as other needs that were identified by the agencies.

A survey of data processing equipment in use by state agencies was also conducted to determine what type and how much equipment would be needed in the agencies to make a statewide system operational. The cost of the required equipment was then determined based on the results of the survey.



CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

The next step in the process was a presentation to the Information Management Commission (IMC) to gain approval of the planned enhancements to the system and equipment requested. This committee monitors the data processing and information processing systems and equipment purchased and operating within the state to insure they conform to the standards set by GCCC. The request for the dollars received tentative approval and looks very positive for final funding by the Joint Budget Committee of the legislature. The actual equipment purchase and design work is scheduled to begin on July 1, 1990.

To assist with the system design work and to meet the needs of the agency users, a users group has also been established. It is made up of representatives from state agencies and the Department of Personnel.

- PROJECT 4: DEVELOPMENT OF UNIQUE TESTS

Colorado is partially decentralized for exam administration activities. One activity retained solely by central personnel is the development of new written objective tests. The Personnel Department provides written objective tests for agency users and also provides a test bank of other types of previously used exams: structured orals, panel assessments, narrative training and experience evaluations (T&Es), performance items, checklist T&Es and other non-written objective tests. These types of test materials are loaned to agency staff upon request.

As one part of the new selection process, a project was initiated to plan for the development of new written objective tests for specific content areas. (New development of other types of tests will be deferred until later.) The project team developed a plan for a needs assessment survey of all agency personnel staff who administer tests. Following the completion of the user survey, this team will analyze usage data and agency requests for new or revised tests, with the goal of compiling a list of exam development needs ranked in priority order. This priority list will be used as a guideline for the test research unit in its production efforts.

This team will also work closely with the job analysis project to determine the category labels and factors to be used in coding items for an automated item banking system being planned by another team. All items presently in the tests within the exam bank will be coded and entered into the item bank. All newly developed test items will also be coded for entry into the item bank. Ultimately, the item bank software will be capable of storage and retrieval of multiple choice, true-false and essay type items.



CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

Another consideration of this team was the identification of resources to assist in the new development process. As a part of the needs assessment, agency staff are being asked whether they are willing to work as part of the development team, or whether they might (if technically qualified) be interested in undertaking responsibility for new development themselves.

This project has finished its survey design and pilot testing stages and is now in the process of conducting the site interviews to complete the questionnaire. The 7 team members will interview 27 agency staff persons to complete the project. Data analysis will occur in May, with a target date of December for completion of the priority list and coding recommendations.

- PROJECT 5: ITEM BANKING

The Department of Personnel currently maintains a bank of multiple choice tests. Many of the tests contained in the test bank are dated and used to test a number of different job classes. Limited resources in the department restrict the development of new and unique tests. With the development of an automated item banking system it becomes relatively easy to assemble a unique combination of items from the item bank.

The Item Banking team was formed to examine the available item banking software available and determine which would be the best to meet the objective of the project, to provide a central automated test item bank to allow access to all items for exam development. A set of function requirements was developed to be used in the evaluation of the software packages available. Some of the functions included were: 1) production of a camera-ready test containing items selected according to user-defined criteria; 2) storage of item information such as item difficulty and item usage (made available when a user produces a test); 3) the software randomly selects items within user-defined categories or the user selects items manually using pre-defined criteria; 4) items are easy to edit and maintain.

A mail survey of other state governments was conducted to determine if anyone else was currently using an automated item bank, what software/hardware they were using and how it was working. Follow-up phone calls were made to respondents with item banking to gather detailed information and assist in evaluation of the software.

After evaluating the available software, the project team agreed that the development of in-house item banking software would be the best solution to meet the needs specified. Currently a proposal has been sent to the Computer Systems Unit requesting programming assistance to develop and program an item banking system.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

Two additional programs which have been initiated in the last several years but are not directly linked to New Directions have had high peripheral impacts on this initiative. These are the Personnel Certification Program and the Internship Training Program. They are briefly described below:

- THE PERSONNEL CERTIFICATION PROGRAM (PCP)

Since 1985, the Colorado State Personnel System has maintained a training program for personnelists who work for the State. It covers classification principles and procedures, personnel evaluation, affirmative action, personnel rules and the principles and procedures of personnel selection.

The classes that train for personnel selection were divided into seven parts, covering different aspects of personnel selection. The implementation of New Directions has necessitated the addition of an eighth class. The new class in PCP covers a more thorough method of job analysis than has been the case in the Colorado system to date.

The New Directions job analysis method, as it has been called, is more thorough in that it fulfills all known legal requirements of a job analysis. More importantly for New Directions, it results in a "profile" of KSA's (Knowledge, Skills and Abilities) indicating the level needed at hire of each important KSA for the job being analyzed. (See Project 1 - Job Analyses and KSA Profiles.)

The goal of the new class is to train state personnelists in the New Directions job analysis method so that all future job analyses will be uniform and complete. The final product is seen as a complete job analysis bank representing every job in the State system.

- PERSONNEL INTERNSHIP TRAINING PROGRAM

The State Classified Personnel System is a complex system which requires both theoretical knowledge and practical application in the technical aspects of public personnel administration. The Personnel Certification Program (PCP) provides some theoretical knowledge needed in Classification, Selection, Affirmative Action, Performance Appraisal and Rules Interpretation. Completion of this program as indicated by successfully passing the various certification tests is required of all personnelists. Before 1988 there was not a formal training program for new personnelists to learn the practical aspects of public personnel administration. Both agency personnel administrators and the Department of Personnel staff recognized that an internship training program for teaching the practical aspects of personnel was necessary.

CONFERENCE PROCEEDINGS  
Assessing Workforce 2000

The internship training program was designed as a "hands on" program for new personnelists to learn the practical aspects of the various technical areas within the State Classified Personnel System. This training is under the tutelage of qualified experienced personnelists in each technical area (e.g., selection, classification, etc.). The training in each technical area is a separate module. New personnelists work in the Department of Personnel in the assigned technical area for a specified time to complete specific tasks. As the specific tasks or assignments are completed by the intern, they are reviewed and evaluated by the trainer. The intern has to complete the assigned evaluation activities successfully in each area before they are allowed to do the work independently. The length of the internship differs based upon agency needs, intern capabilities and the amount of time the intern spends in the Department of Personnel during the internship time (i.e., 1 day/week vs. 2+days/week). It is estimated that the internship takes approximately 3 months.

Evaluation of the intern is done throughout the training period. Additionally, there is an evaluation of the intern upon the completion of the internship. This evaluation and feedback is given to the intern as well as to the interns's supervisor. It is recommended that an additional evaluation be done approximately 6 months after the completion of the internship by the supervisor. If deficiencies are noted, further training could be recommended.

Much remains to be done but most projects are on schedule and some are ahead of schedule. The course has not been completely smooth; changes have occurred in project activities and teams have needed to remain very flexible in their planning. After eighteen months of New Directions in Selection, there is optimism that system change will occur as planned and that applicants, managers and taxpayers will benefit from the many improvements in the selection process.

## AUTOMATION OF TENNESSEE'S EMPLOYMENT SYSTEM

by: Robert Perry

Tennessee is very similar to most states in the way we test and score applicants for State employment. The Executive Branch of Tennessee State government is approximately 90% civil service and covers some 37,000 employees in 1,250 different job classifications. With only minor exceptions applicants may apply for any of our positions on a day-in day-out basis. Program examinations in Tennessee are thus an exception rather than a norm, this applies both to T & E examinations as well as written and/or performance tested job classifications.

Our system of score reporting, placement of applicants on employment registers, working of registers and the actual appointment of new employees, again like most States, is fully automated. Applicants must however identify specific State job titles of interest, submit an application for employment and be approved or rejected for written tested classes by our interviewing staff or be scored or rejected for T & E rated classes by our rating staff. Once an applicant has a score and they wish to increase their score, they must reapply and be reinstated or reevaluated.

What is fairly unique about Tennessee's system is that approximately 2 1/2 years ago we implemented an on-line scoring system for our T & E rating system and began capturing and storing an applicant's education and work experiences in a variety of data base formats.

Our goal was to speed up our rating process establishing a turnaround time for T & E ratings of approximately one week and one day for written and performance tests. Our volume of applicants is approximately 35-40 thousand different applicants per year applying for approximately 120,000 job classifications per year. We also wanted to simplify the reapplication process for current employees and establish an employment history computer record for non-State employees to capture falsified applicant records. State employees are not required to submit a complete application each time they apply, only an update, while non-State employees must still submit a complete application but their prior application data is immediately accessible for verification.

An applicant's education and experience data is coded and entered into our system using an education coding scheme for educational level and academic majors based on the National Center for Education Statistics "Classification of Instructional Programs" which includes approximately 1,200 different majors. Additional codes were however, developed by my staff to meet specific classification requirements such as State licensure or certification. Coding of experience is based on a modified Dictionary of Occupational Titles, DOT, and Standard Occupational Classification, SOC, coding scheme which defines the occupational group, specific job description, job level (whether trainee, worker, lead worker, supervisor, etc.) and finally EEO code. Additional information including number of months of employment, date of last employment and date degree was conferred is also collected.

Just the capturing of such data in our current system has led to numerous special studies, i.e., employee literacy, hiring trends, work force analysis, etc. However our current system was designed to capture and store such data for greater improvements in our ability to match applicants to jobs and jobs to applicants, and is basically oriented towards career development of employees and job placement of applicant.

Design activities are currently underway to establish a position profile for each job in State government and a comparable profile of applicant vocational interests which can be used to conduct preliminary applicant screening for our 1,250 different job classifications.

When developing our profile questionnaire a variety of vocational or occupational interest checklists were reviewed, i.e., Ohio Vocational Interest survey, Gordon Occupational Checklist, Jackson Vocational Interest survey, Strong Vocational Interest blank and the Minnesota Importance Questionnaire. Of these the Minnesota Importance Questionnaire, MIQ, which is based on the theory of Work Adjustment recognizes both worker needs, i.e., pay, co-workers, supervisors and job reinforcers, job conditions that meets these worker needs, came closest to being the type of instrument that we were looking for. Associated questionnaires including the "Minnesota Job Description Questionnaire" and the "Minnesota Occupational Classification System" creates job profiles based on required job abilities, i.e., perceptual, cognitive, motor, etc., and values, i.e., internal, social, environmental, in a manner similar to the "Minnesota Importance Questionnaire" which is directed toward the applicant's needs. However, for our specific purpose and our particular system we decided to opt for a most specific approach to capturing job interest data.

One of our biggest problems in working registers and making employment decisions in Tennessee has nothing to do with finding qualified applicants but involves contacting, interviewing, and working our way through a variety of applicants for each position vacancy who are not interested because of salary, location, work shift, work condition, environment and numerous other conditions. Other factors having to do with specific job responsibilities which have skill requirements such as reasoning ability, mathematics and language or communications are not communicated to the applicant at the time of application. To be perfectly frank most applicants have no idea what our jobs might consist of, might pay (actual salary that can be offered) and might require as far as job interests or skills until they are contacted regarding an interview.

From our perspective what was lacking was better communication to the applicant about job conditions, i.e., pay, location, hazards, work conditions, etc. Second, information from the applicant concerning their vocational interest, what type of jobs are they interested in performing. Applicant or self assessment of skills and their ability and interest in performing certain types of tasks on a regular, routine basis. Finally, being able to quantify this information both from an applicant and position basis to permit comparisons and document our actions.

The applicant job profile questionnaire which was developed and is presently used as a counseling tool is being integrated into the design of our automated applicant scoring system. The system which we are currently designing and programming requires that the applicant complete a standard employment application, if the applicant's education and experience is not on file, and a job profile questionnaire, no job titles are necessary.

The applicant responses or profile are run against our class/position profile data base. When the applicant's profile matches at least one position in a given class, a temporary applicant class/experience file is created. If the applicant's profile does not match any position for a given class this information is written directly to the applicant's record. For every class/applicant match the applicant's education and experience is evaluated by the computer to determine if the applicant meets any of a variety of education and experience



substitution requirements necessary to qualify for scoring. If the applicant does not qualify for a given class a record is again written directly to the applicant record. If the applicant does qualify the computer completes scoring of the applicant education and work experience. Storage of experience/education codes, points and formulas require approximately 450 megabytes. Obviously storage of applicant education, experience, classifications, classification status, profile, etc., requires considerably more.

The system once on-line will reduce and/or "eliminate" classification rejection complaints, improve our applicant pool, establish true career development opportunities and reduce overall complaints.

What I have discussed thus far are general terms about how you apply for employment consideration and the reporting of scores to applicants for classifications in State service. Employment opportunities however do not actually occur until vacancies exist and efforts are made by an agency to fill a particular vacancy.

When requesting a list of candidates to fill a vacancy the agency will enter the class, position and the position's profile into their terminal. Applicants with scores for a given classification will be extracted from the data base, their profiles matched to the particular vacancy in question and if a match occurs, rescored. This scoring insures that employees always have up-to-date scores when a vacancy occurs. The applicant is rescored only for the job classification in question and only if they are a State employee. Rescoring of employees is possible due to up-to-date employment information, non-State employees can be rescored only with reapplication and the updating of their education/experience records.

Applicant rescoring results are in turn reported to the applicant along with their standing on the individual register.

This completes my summary of both our existing and planned applicant system. Given our current schedule, full implementation of the revised scoring system will not be accomplished until 1995.



# LIST OF ITEM DESCRIPTIONS

## EXPERIENCE UPDATE SERIES

10-10-00

SSN 1 1210001 0000 0000 00 000000 EXP 0 00000000 0 00000000 00000000  
 HIR CODE 0000 0000 00 000000 EXP 0 00000000 0 00000000 00000000  
 00000000 00000000

ACTION		DESCRIPTION		HIR CODE		HIR CODE	
				EXP		EXP	
1.	1210 0000 00	CHIEF PROGRAM DIR.	10	00	00	00	00
2.	2010 0000 10	CHIEF MEMBER 10	10	00	00	00	00
3.	2010 0000 10	CHIEF MEMBER 10	10	00	00	00	00
4.	1210 0000 00	CHIEF MEMBER 10	10	00	00	00	00
5.	1210 0000 00	ADMIN ASSI 0	00	00	00	00	00
6.	1210 0000 00	CHIEF MEMBER 10	10	00	00	00	00
7.	2010 0000 10	CHIEF MEMBER 10	10	00	00	00	00
8.	2210 0000 00	CHIEF MEMBER 10	10	00	00	00	00
9.	2010 0000 00	CHIEF MEMBER 10	10	00	00	00	00
10.	2010 0000 10	CHIEF MEMBER 10	10	00	00	00	00
11.	2010 0000 10	CHIEF MEMBER 10	10	00	00	00	00
12.	2210 0000 00	CHIEF MEMBER 10	10	00	00	00	00
13.							
14.							
15.							

EXP 1210 0000 00 000000 0000 0000

OPT 1210 0000 00 000000 0000 0000

STATE OF TENN APPLICANT SYSTEM

EDUCATION UPDATE SCREEN

04/19/96

SSN 954 73 8783

NAME

BOHDEE

BUGNE

EDUCATION

1990 199 YEAR MASTERS DEGREE

DATE COMPLETED

04/96

CODE	ADDED	CODE	COURSEWORK	HRS	ADDED
199	01-90	1	287 SOCIAL WORK	000	01-90
354		2	454 SOCIAL SCIENCES CLERICAL	999	01-90
		3	000	0	
96	01-90	4	000	0	
117		5	000	0	
		6	000	0	
0		7	000	0	
000		8	000	0	
		9	000	0	
10	00-00	10	000	0	
		11	000	0	
0		12	000	0	
		13	000	0	
		14	000	0	
		15	000	0	

XXX LISTED DATA FOR DELETED

TOP 199 954 73 8783

## STATE OF TENN. APPLICANT SYSTEM

## APPLICANT INQUIRY SCREEN

03/19/90

PAGE 001

SSN: 451-24-8783 NAME: DODGE DUANE A  
 DT APPL: 02-23-90 ADDR: 4407 CLINDALE SQ  
 DELETE FLAG: NO CITY: NASHVILLE TN 37204-0000 B. PHONE: 615-241-3211  
 BUDGET CODE: 33905 NET PTS: 0 BONUS PTS: 4 LEGAL CL: CAMPBELL  
 DEPT PHONE ONLY: NO US CITIZEN: YES CC PREF: 12 94 00 00 00

## CLASSES AND SCORES

CLASS NO	CLASS ABBREV	TEST	EC	WRT	PRF	EXP	EDU	SHS	REF	RES DATE	S	R
071572	PCY TCH CO E	53051	015	000	000	000	000	000	000	01-28-90	9	
072232	MR PRG SPC 2	53679	019	000	000	082	082	000	000	02-06-90	0	
072237	MR PRG SPC 3	53680	019	000	000	083	082	000	000	01-28-90	0	
078841	LIA TCH COUP	53042	019	000	000	073	072	000	000	02-01-90	0	
079661	PSY SOC WL 1	53614	019	000	000	085	084	000	000	02-29-90	0	
079662	PSY SOC WL 2	53615	019	000	000	080	084	000	000	03-05-90	0	
079693	HS PRG SPC 3	53625	019	000	000	085	084	000	000	01-29-90	0	
079694	HS PRG MGR	53626	019	000	000	085	084	000	000	01-29-90	0	
079851	HS PRG DIR 1	53631	019	000	000	085	084	000	000	01-29-90	0	

OPTION: TRY: 661 REF: 451245293

0-001

# UNITED STATES DEPARTMENT OF JUSTICE

## APPLICANT UPDATE SCREEN

01-19-90

SSN: 451-24-5221 NAME: DODGE, JAMES E. DOB: 01-19-40

ADDR: 4407 GILBERT RD CITY: BIRMINGHAM ST: AL ZIP: 35204-0001  
 0000000000

HOME PHONE: 205-251-7111 OFFICE PHONE: 205-241-7111  
 0000000000 0000000000

EXM CTRF: 12 LPRAL CTRF: 12 US CITIZEN: N  
 00 00

COURT FEE CTRF: 12 94 00 00 00  
 00 00 00 00 00

NEW SSN: 0000-0000-  
 DEPT FROMO ONLY: 10 DISBURSE: 97-07-46 RACE: 0000-0000  
 00 00 00

\*\*\* ENTER ALL FEES DEPOSIT \*\*\*  
 OPTION: 100: 0000 100: 0000

01-19-90

## The High-Low Predictive Validity Design: High Power With Small N

By Joel P. Wiesen, Ph.D.

Massachusetts Department of Personnel Administration

14th Annual IPMAAC Conference

June 27, 1990

If what I say today makes sense, and I think it does, it may be technically feasible to conduct predictive validity studies with N's of 20 and 30. I will describe a unique methodology for conducting predictive validity studies for personnel selection which deals with the problem so clearly described by Schmidt, Hunter and Urry (1976). We all know the problem of power and sample size which Schmidt described, but let me describe it briefly. With typical size correlations between predictors and criteria, the usually available sample sizes are just not adequate for research. Power in research is a lot like looking for lost car keys. If I look for 5 minutes I may find them and may not. It depends on how lost they are, how efficiently I search, and how lucky I am. The 5 minutes is analogous to the sample size we have to do the research. For example, let's say that we will do a validation study when the relationship between the test and the criterion is .2. (Of course, in real life we will not know the true correlation, just as we do not know how lost the car keys are, but let's assume we have reason to think the correlation is .2.) Schmidt, Hunter and Urry showed that we would need to conduct our validation study with a sample of 280 to have a 90% chance of finding a significant correlation. In other words, we would need to have a sample of 280 to have a power of .90. Also, this assumes we will hire randomly; otherwise we will need an n larger than 280. The problem is that we typically do not hire 280 people at one time and we often do not hire people randomly, so predictive studies are typically not technically feasible. Even concurrent validity studies may not be possible in many situations where we have fewer than several hundred employees. The new approach I will describe today may make it possible to conduct predictive validation studies in situations where we are hiring only 20 people.

Before anyone who heard that last sentence has a chance to leave, let me put more emphasis on the word MAY.

What I am presenting today is still in its developmental stages. I will present the research I have done to convince myself that this approach is real and not wishful thinking. I will also present some practical guidelines for the use of this technique, but the emphasis will be on presenting the support I have for the existence of a new validation research design. I am now looking for situations to use this approach. I am very interested in talking to and perhaps working with anyone who may try to use it. Please feel free to stop me anytime you see me at this conference to discuss this topic or call or write after the conference. I look to the discussant and the audience to raise questions.

I have three goals for this presentation: To describe the new research design for predictive validation, to show that this research design has greatly improved power, and to provide some practical guidance in the use of this research design.

### The New Research Design For Validation Research

The new validation research design is easy to describe. Rather than hiring the top scoring applicants, and rather than hiring a random selection of applicants, this new design calls for the hiring of some people who scored very high and some who scored very low on the predictor. For example, if you had 1,000 applicants and planned to fill 20 job openings, this design would have you hire the highest scoring 10 applicants and the lowest scoring 10 applicants. (I realize there may be practical limitations to such a pattern of hiring, and I will make some practical recommendations to deal with this later.)

I would like to describe the evidence I have that this high-low design provides much more power than the usual alternatives of hiring randomly or hiring only high scoring applicants. But before we can consider that evidence, we need a statistical test which is appropriate to the data. So first we will look at the statistical test which I think is appropriate and then we will turn to the data on the statistical power of the new research design.

### Method of statistical analysis

To evaluate the power of a research design we need to have a method of statistically analyzing the research data. I propose doing this with the familiar formula and tables of significance which are used to calculate and evaluate the statistical significance of the Pearson correlation coefficient. It is clear that we can use the mathematical formula of the Pearson correlation coefficient to describe the observed relationship between the predictor and the criterion variable. In only one case will this new statistic be the same as Pearson  $r$  which would result from the usual predictive validity research design. But in all cases, this new statistic will have a range from minus one to plus one, and we expect it will equal zero when there is no relationship at all.

The one case where the high-low correlation coefficient will have the same distribution as the Pearson  $r$  is when the null hypothesis is true; that is, when the population correlation is zero. (At least the usual null hypothesis is that the population correlation coefficient is zero.) Since under this null hypothesis the new statistic will have the same distribution as the Pearson  $r$ , the usual various statistical tests of significance, either one-tailed or two-tailed, will be appropriate.

I might mention that although I was convinced by logic that the distribution of the correlation for high-low samples ( $r$  high-low) is identical to the distribution for Pearson  $r$  under



the null hypothesis that  $r$  equals zero, I also had a small fear that there may be an artifact that was not apparent to me which leads to a high value for  $r$  high-low even when the population correlation is zero. So I checked this using the Monte Carlo approach which I will describe shortly. But there was no indication of any artifact: When the population correlation equaled zero,  $r$  high-low was also zero.

But let me be very clear: A major distinction between this new statistic and the Pearson  $r$  arises whenever the population correlation is not zero. For that reason, after we consider the question of power we will consider the shape of the relationship between the new statistic and the familiar Pearson  $r$ .

### Approaches to Determining Power

I will not present an analytical, mathematical analysis of the statistical power of this new research design using this new statistic. I have not found a way to do this.

Absent a mathematical treatment of power, there are two ways to demonstrate the statistical power of this new research design: a hard way and an easy way. The hard way would be to conduct many pairs of research studies; one study of each pair would be done using the usual predictive validation approach and the other study would be done using this new research design. This approach has obvious practical limitations. Instead I chose to use a simulation of research rather than actual research to evaluate this new research design. The simulation involves creating populations of numbers with known correlations and sampling from these populations using the usual and the new approaches. This approach has been referred to in the literature as a Monte Carlo study.

I will report today the results of 10,000 validation studies. All these studies were simulated, but they were simulated using a method which closely approximates what may be found in the real world.

### The Monte Carlo method

I decided to look at a typical predictive validation situation in which there are more applicants than openings, there is a selection test (the predictor) which is used to select some applicants to be hired, and there is a measure of job performance which serves as the criterion. Of course, the measure of job performance is only gathered for those people who are hired.

In this Monte Carlo study, I created many (bivariate-normal) populations with known correlations. The correlations I looked at were chosen to be typical of those which might be seen in real life. I created populations with population correlation equal to either .1, .2, .3, .4 or .5.

Then I arranged to take several different types of samples

from the populations. First, to model the traditional ideal practice in a predicative validation study, I randomly selected from a population. (Although true random sampling is rarely done in predictive validation research, it is the ideal approach.) Second, to model the more usual (but inefficient) practice in predictive validation research I sampled only the top scoring applicants from a population. Third, to model the application of the new research design, I sampled both the highest and the lowest scoring applicants.

This was repeated many times. Typically I ran 100 replications. For each replication I started from scratch, generating a new population with known parameters and sampling from that population.

### Evidence of Higher Power

Now let's look at some of the results of the Monte Carlo research starting with a likely typical situation with a correlation of .20 in the population. Now the underlying correlation may be higher, but let's assume that the nature of the test and the job, and the unreliability of the criterion and the predictor, has led to .20 being the actual observable correlation between test and criterion in the population.

DISCUSS FIGURES 1 to 3 HERE (showing power under each of 3 different sampling schemes).

Let me summarize the improvement in statistical power seen in these figures using our original example with a population correlation equal to .20. In this case Schmidt, Hunter and Urry would suggest that we need a sample of 280 to have a power of .90. Using the high-low research design, if you have about 2,000 applicants and are willing to hire a group of 20 people who include the 10 highest and the 10 lowest scoring applicants, you will have power equal to .90.

Now many jurisdictions do not hire 280 people at a time, but many of these same jurisdictions do have 2,000 applicants and many do hire 20 people at a time. So an important, and even crucial, type of research which was once thought to be impossible due to technical limitations is feasible using this research design.

### Relationship of Pearson $r$ to $r$ high-low

If the observed  $r$  (let's call it the Wiesen  $r$ ) is statistically significant, we may wish to interpret it in terms of the Pearson  $r$  which we would expect to find if we use a more traditional random sampling design. The Monte Carlo studies show a smooth, monotonic relationship between Pearson  $r$  and Wiesen  $r$ . This relationship seems to depend on the population size, the sample size, and the population correlation coefficient. We can use this observed relationship to shrink

the observed Wiesen  $r$  back to the actual population Pearson  $r$ .

DISCUSS FIGURES 4-5 HERE (showing the relationship between Pearson  $r$  and  $r$  high-low for  $r=.2$  for two population sizes.)

### Practical Considerations

There are two types of practical considerations: those arising from the fact that this technique has not been fully explored, and those arising from the nature of the selection process required by this technique. For today, we will not talk further about the limitations posed by the admittedly incomplete analysis of the distribution of the test statistic or related matters. Instead, we will focus on the questions of practical implementation.

The public sector should often be able to meet the one major requirement of this new research design: large number of applicants. Even when we have only a few openings, we may get a large number of applicants.

One major practical difficulty is convincing an organization to hire the lowest scoring applicants. One possible way around this is to change the design a bit and instead of using the applicants from the top and bottom of the distribution, use the applicants from the top and the middle of the distribution. I think it will be possible to sell this less extreme approach to employers. We can call this the high-median approach. I have run some Monte Carlo simulations of this and it looks promising.

DISCUSS FIGURE 6 HERE (showing the mean observed  $r$  under 4 different sampling schemes, including  $r$  high-median).

We may be at the brink of a new age in validation research. Just as new technology has enabled almost everyone to own a computer, it may be that this new technology will enable almost everyone to run predictive validation studies. Perhaps now jurisdictions of all sizes will have to make decisions which until now were made only by the largest jurisdictions, namely: Should we run a predictive validation study? Should we sample from the top or use a high-median design?

### Literature Review

The high-low design for correlational research is virtually unknown in the literature and the high-median design seems to be totally unknown. However, this selective sampling approach is closely related to traditional experimental design. In traditional experimental design, when we are trying to see the effect of an independent variable on a dependent variable, we choose levels of the independent variable which are extreme. If we want to see if the relationship is linear, we choose a third

group with a level of the independent variable which is mid-way between these two levels.

This salutary effect of restriction in range to the extremes has not been totally ignored with respect to correlational research. For example, Snedecor & Cochran (1967) give one example of this in their text on statistical methods. In the example they take the highest 5 and lowest 6 scores and find a correlation of .89 as opposed to .77 in the original sample. (I presume their whole sample was small.) However, they do not suggest this approach as an experimental design, but rather present this example to show the effect of selective sampling.

Peter and Van Voorhis (1940) describe a procedure similar to mine, but different in at least two respects. First, they use the point biserial correlation coefficient rather than the Pearson  $r$ . Second, they do not recommend the approach as a way to improve statistical power. They only present the approach for use in those situations where only the extremes of the whole population are available for study. They give a quite sophisticated mathematical analysis of the problem and I have just located a copy of the reference. So I can not say too much more about it, other than to say they have treated some aspects of the problem and not other aspects. After a quick reading of their presentation I now think it may be possible to approach the question of the distribution of high-low and high-median correlation coefficients mathematically rather than by using Monte Carlo simulation.

So I have found a few, old references which discuss correlational approaches such as those proposed today. There is one detailed treatment in the literature of an approach which is close to the  $r$  high-low, but so far I have found nothing in the literature similar to the  $r$  high-median. Nonetheless, the logic underlying these approaches is firmly within the tradition of experimental design.

### Conclusion

I find this research very exciting because I think it has great promise. But, originally, I had reservations. I am not a mathematical statistician. I was concerned that I may be missing something or fooling myself with incorrect assumptions. So I did a number of extra checks and took some extra precautions in doing this Monte Carlo research. I already mentioned the check for some spurious factor which could result in high  $r$  high-low even with a population correlation equal to zero. I was also concerned that the program which I used to generate the random numbers may have been faulty. Typical random number generators are not truly random. Sometimes random number generators can get into loops and generate numbers which are very non-random. I found no evidence of this with the random number generator I used, but nevertheless I used a conservative method to conduct the simulations. When choosing random samples of size 20, the simple approach would have been to assume that the random number generator worked perfectly and

just use it to create a parent population and to repeatedly take samples of 20 cases from that parent population. Rather than do this, I repeatedly generated a parent population and then sampled once from each parent population. This was a bit less elegant (and much more work), but I thought it would avoid a possible weakness. I mentioned that I was concerned that some unknown spurious artifact might be yielding my high observed values for  $r$  high-low. So I ran a few simulations of the high-low variety with the population rho equal to zero. In those cases,  $r$  high-low was also zero. Finally, early in this research I ran a few simulations with  $z$  score cutoffs and compared the results with the values in Schmidt, Hunter and Urry (1976) and found my values agreed closely with theirs. So I am now rather sure that the Monte Carlo method I used did function as intended.

So although originally I was not convinced the effect was real or that it was a contribution to the literature, now I am sure it is a real effect and I am starting to be sure that it is a contribution to the literature. I await the comments of Barbara Showers, our discussant, with great interest.

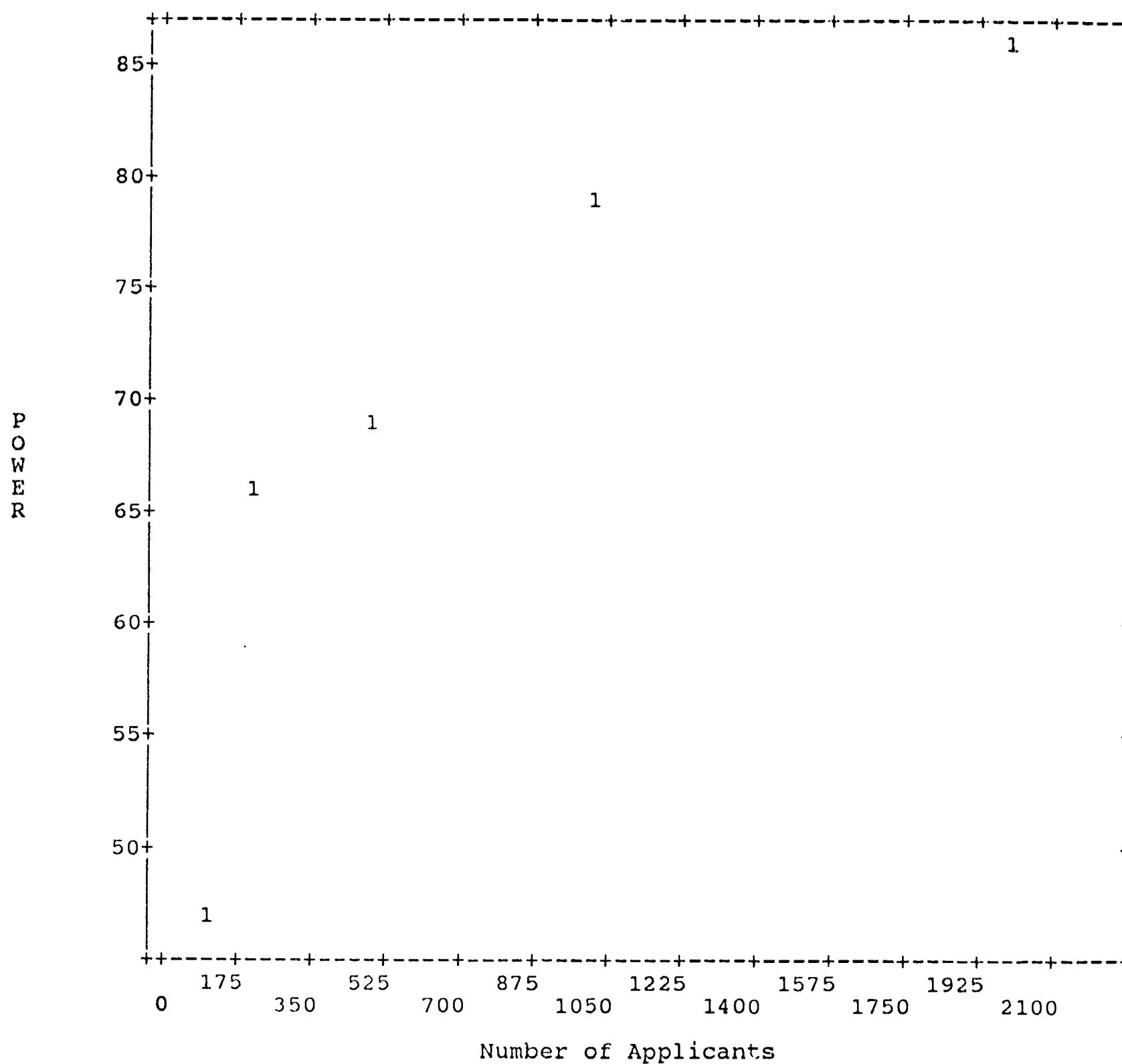
As I collect more and more data I grow more and more convinced that this is indeed an important experimental design for correlation studies. It allows predictive validity studies to be done when otherwise none would be possible due to small  $N$ . The next step is to try it out in a real-life setting. If you have a situation where you think this approach may be fruitful, please let me know.

#### References

- Peters, C. C. & Van Voorhis, W. R. (1940). Statistical Procedures and their Mathematical Bases. New York: McGraw-Hill.
- Schmidt, F. L., Hunter, J. E. & Urry, V. W. (1976). Statistical Power in Criterion-Related Validation Studies. Journal of Applied Psychology, 61, 473-483.
- Snedecor, G. W. & Cochran, W. G. (1967). Statistical Methods (6th ed.) Ames, Iowa: Iowa State University Press.



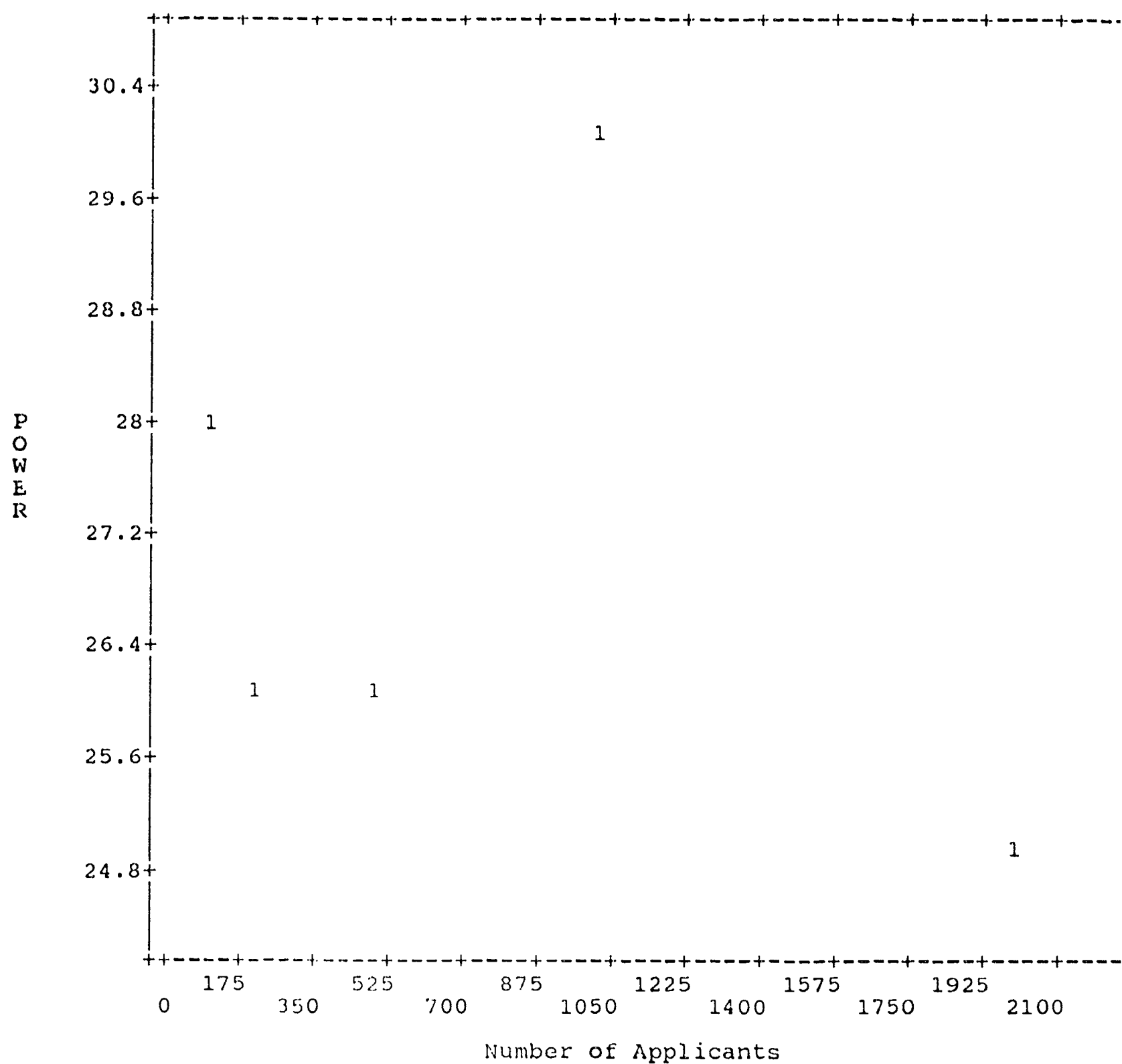
Figure 1. Power curves for high-low sample  
(Population Correlation = .2)



5 cases plotted.

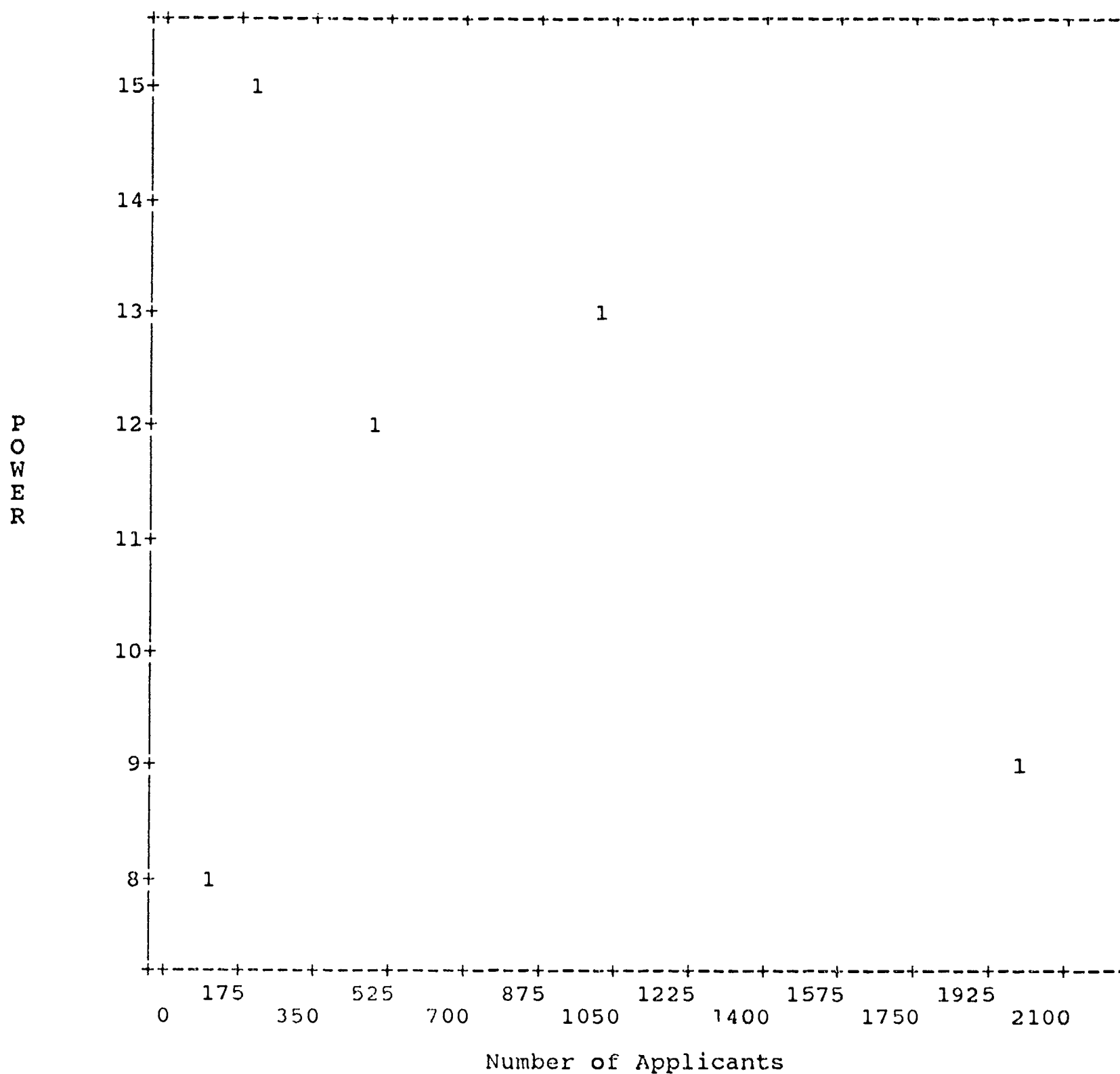


Figure 2. Power curves for random sample  
(Population Correlation = .2)



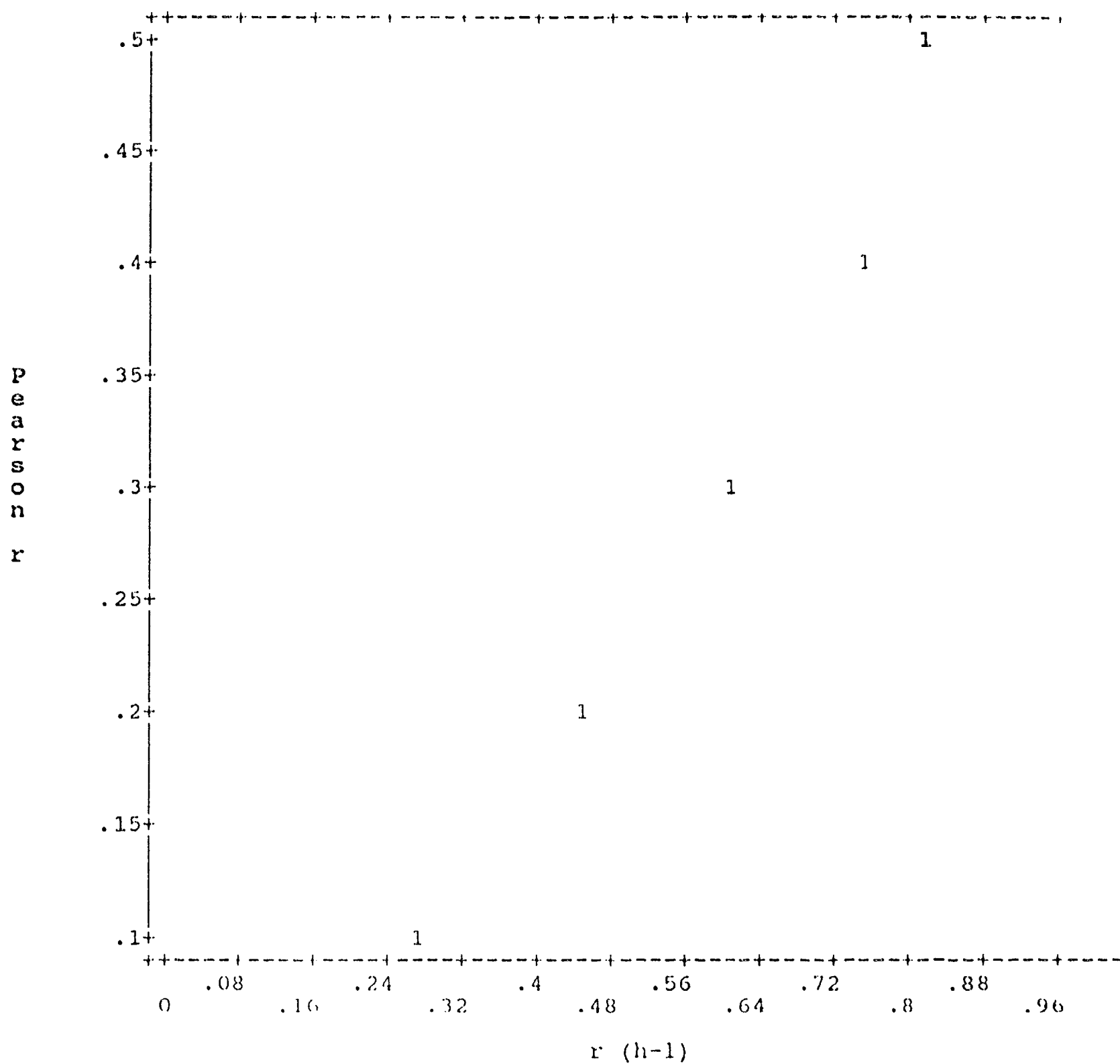
5 cases plotted.

Figure 3. Power curves for top sample  
(Population Correlation = .2)



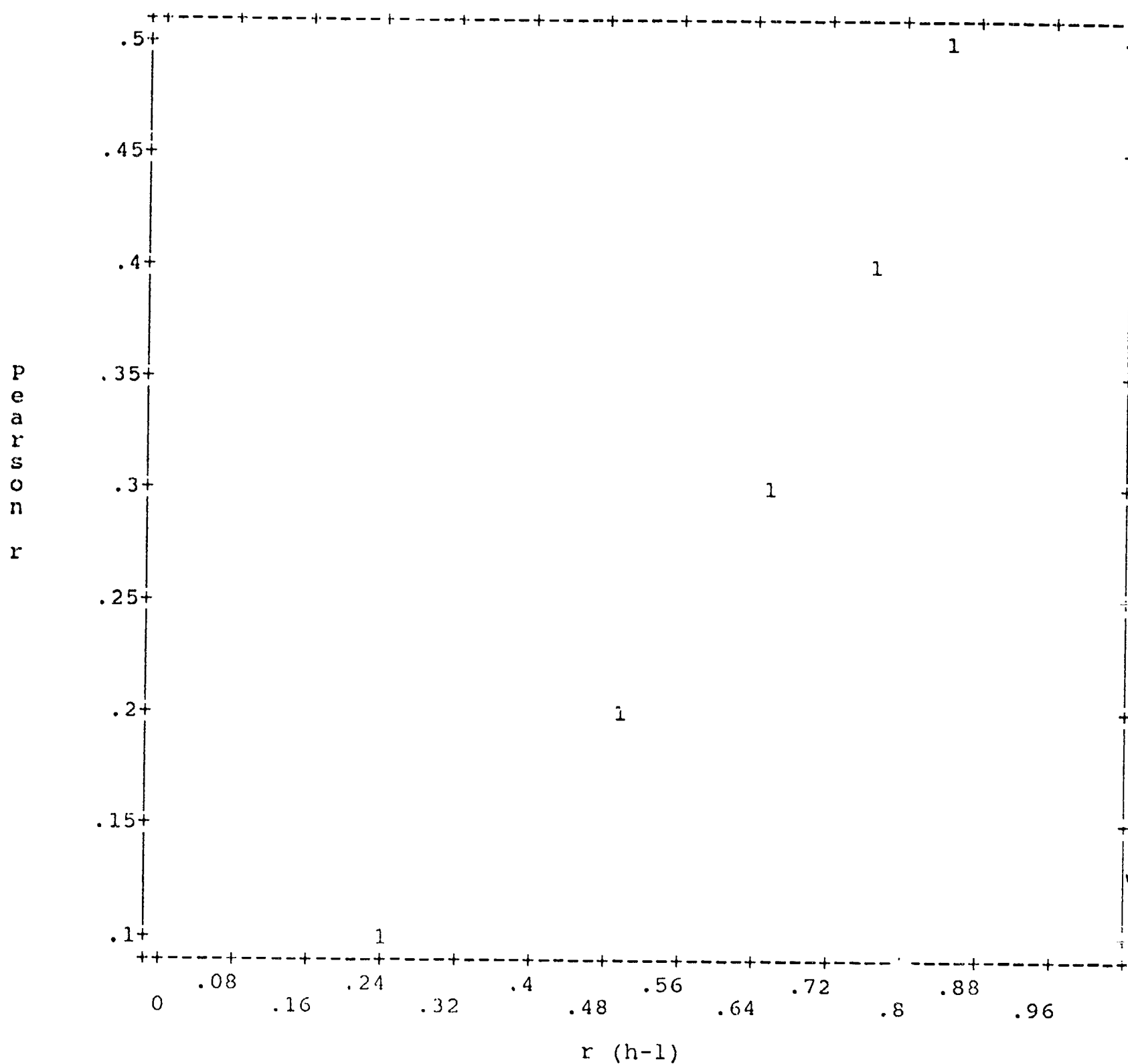
5 cases plotted.

Figure 4. Relationship between Pearson  $r$  and  $r$  high-low  
(Pop. Corr.=2, sample = 20, no. of applicants = 500)



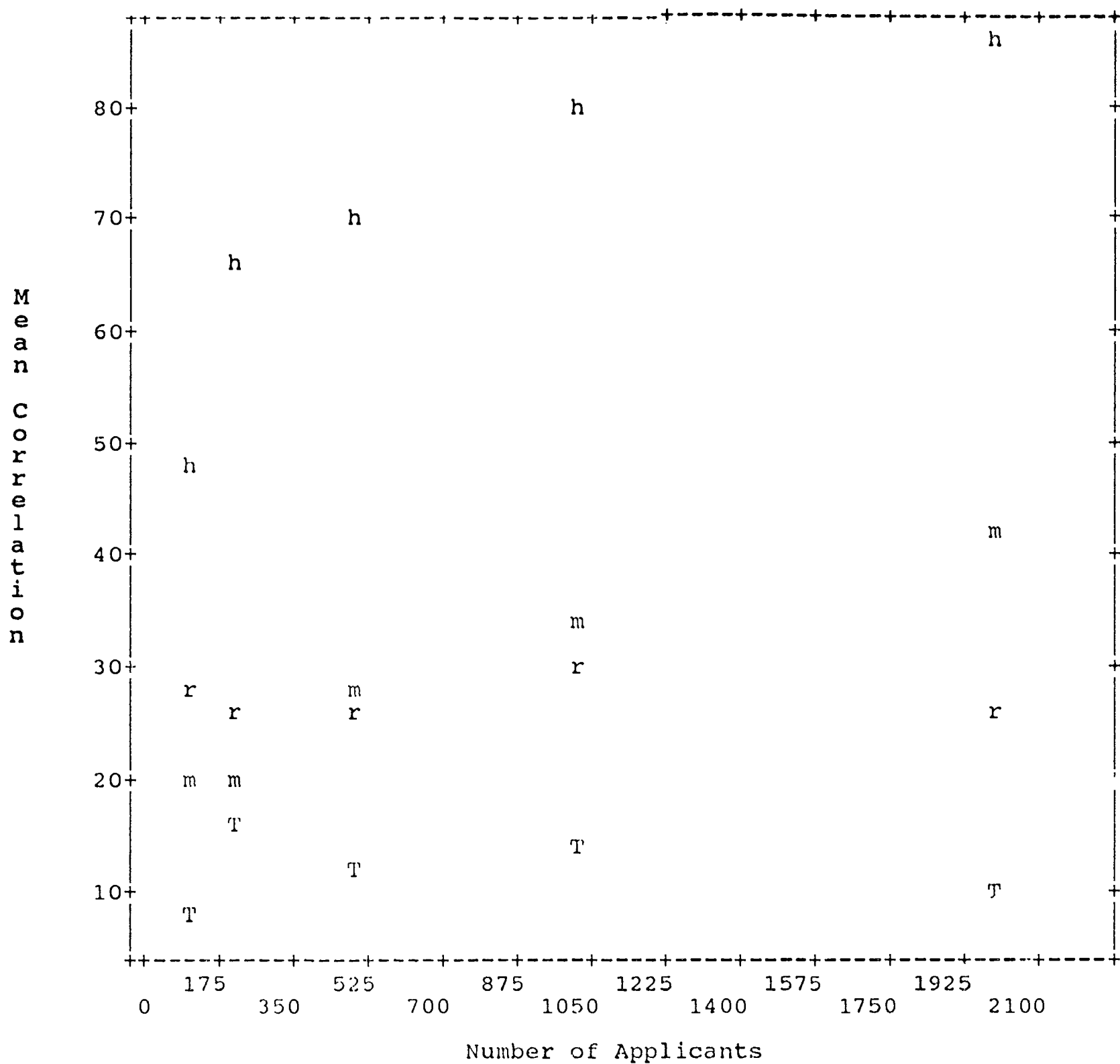
5 cases plotted.

Figure 5. Relationship between Pearson  $r$  and  $r$  high-low  
(Pop. Corr.=2, sample = 20, no. of applicants = 1000)



5 cases plotted.

Figure 6. Mean observed correlation for four types of samples  
(Population Correlation = .2, sample = 20)



Key:  
h = high-low sample  
m = high-median sample  
r = random sample  
t = top sample

"POWER TO THE USERS -- DECENTRALIZING AN AUTOMATED  
EVALUATION SYSTEM AND NOT REGRETTING IT"

PANEL BY

GRANT GILFEATHER  
PERSONNEL ANALYST IV  
ARIZONA DEPARTMENT OF PUBLIC SAFETY

ROB ROCKENBAUGH  
EDP SYSTEMS ANALYST II  
ARIZONA DEPARTMENT OF PUBLIC SAFETY

PRESENTED FOR THE  
INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION  
ASSESSMENT COUNCIL CONFERENCE  
ON ASSESSING WORKFORCE 2000

JUNE 27, 1990



## "POWER TO THE USERS" -- DECENTRALIZING AN AUTOMATED EVALUATION SYSTEM AND NOT REGRETTING IT"

### PROBLEM

How to rescue a centralized automated performance evaluation system run on an IBM-AT from inertia, inflexibility, and inaccuracy.

### HISTORY

In mid-1985, the Human Resources Section of the Arizona Department of Public Safety initiated the first computerized evaluation system for its 1600 employees. This system was run from one IBM-AT personal computer. The software was programmed and maintained by an outside vendor. Over time, as exceptions to the system increased and money available to maintain the system decreased, it became evident the Department's Data Processing Division would be a logical choice to reprogram and maintain the system. Problems with the IBM-AT system surfaced regularly, including ineffective monitoring and reporting of missing evaluations, slow turnaround on processing evaluations involving status changes, and an inherent software code error that deleted employee evaluations files without warning. The problem with ineffective monitoring of missing evaluations caused our merit board, the Arizona Law Enforcement Merit System Council, to cancel the evaluation portion (15%) of a Sergeant's promotional process. With these considerations in mind, the management of Human Resources, Data Processing Division, and the Law Enforcement Merit System Council (LEMSC) decided in June, 1989, to convert from the IBM-AT evaluation system to an IBM 3081 mainframe evaluation system.

### THE CONVERSION PROCESS

#### Requirements Defined

The Data Processing Division management created a "new systems" team to help Human Resources convert its evaluation system to the mainframe. The team consisted of a newly promoted systems analyst and a programmer who had been with the department just over a year. Immediately, the team contacted those in our section responsible for the IBM-AT driven evaluation system and obtained the technical documentation manual left by the outside vendor. Jointly, the system requirements emerged to build an evaluation system that possessed the following features:

- ° a system programmed in one language
- ° timely processing capabilities through multiple input workstations and 24 hour mainframe printer.
- ° an internal on-going system documentation and maintenance program.

- distributive processing for statewide input and output.
- extensive error checking capabilities for location codes, employee status changes, evaluation dates and rater and operator errors.
- regular monitoring and reporting of missing, pending or erroneous evaluations.
- multi-level password, terminal and location code security for keyboard operators and backups.
- regular ad hoc and routine report generation capabilities.
- extensive storage capacity to keep evaluation histories back to 1985.
- compatible on-line access to current mainframe employee personnel data.

#### Problems Solved Through Consultation

Almost daily meetings occurred for about three weeks as the team studied the technical documentation, the existing Employee Performance Appraisal Manual and the Arizona Law Enforcement Merit System Council Regulations and asked questions of Human Resources Section personnel. The team also reviewed key files and programs from the IBM-AT driven system to better understand any file history inconsistencies. Questions raised regarding evaluation policies that appeared inconsistent with the manual, regulations or current practice were resolved through consultation with the Associate Business Manager of LEMSC.

#### Evaluation Rating Scales, Factors and Weights

The IBM-AT driven evaluation system was based on a 12 point rating scale. Ratings 7 and above are standard, while ratings below 7 are unsatisfactory. Employees generally are rated twice a year. The system contained 40 performance factors; three of which are labeled Special/Other and thus customized, and each performance factor was weighted one, two or four by the supervisor. This applied to all the performance factors except the first three; Work Habits, Relationships with People and Policy and Procedures, which were weighted 4. Employees were rated on these three factors regardless of position or rank. Prior to marking the SCANTRON Employee Evaluation Rating Sheet, the rater prepared an Employee Performance Report Documentation Sheet for each performance factor rated. Upon completion of this sheet, the rater filled in a SCANTRON Employee Appraisal Sheet and sent it to the Human Resources Section to process.

### New CRT Screen Layout

The team's approach was to keep the basic layout of the performance evaluation form intact and adjust it to fit on a CRT screen. The team researched two previous years of evaluations and determined that 95% of the raters used twelve or fewer performance factors. Thus, the new CRT screen was designed for twelve performance factors which included a Special/Other factor. If a second screen was required, the operator would press a PF10 key and it would appear. The second screen is also used to extend or remove an employee on probation or mark an interim rating.

### Distributive Processing

It was further agreed by the team and Human Resources management that the input of the employee evaluations be done off another form and by keyboard operators and their backups throughout the department at assigned CRTs. At first the number of operators envisioned was less than 15. The number grew to 66 as various bureaus gave Human Resources management feedback. Following this decision, keyboard operators and backups were identified by the five bureaus and were given assigned terminals and passwords by Human Resources.

### Manual Writing

While the team was coding the new program, Human Resources Section management rewrote sections of the Employee Performance Appraisal Manual and redesigned the new employee performance input form. Key Human Resources personnel wrote the Operator's Manual for the new system.

### Testing

The team continued to reprogram the system from dBase III+ to Natural 1.2. As various segments were completed, key Human Resources personnel were asked to "test" the system and report any problems. Selected Human Resources personnel were also responsible for entering passwords, terminal identification and location codes for keyboard operators and their backups. The security module is only available to Human Resources Section.

### Training

With testing completed, the keyboard operators and backups were trained in a three hour session using six CRTs similar to their work stations. Since DPS is a statewide organization, it took three days and six sessions to train all operators and backups. The team created test data for the classes to use. The new operator's manuals and input appraisal forms were explained. The revised Employee Performance Appraisal Manuals were distributed. The Data Processing Team and Human Resources Section personnel both shared in teaching.

## Security

During these training sessions, the importance of security was stressed. Operators and backups were urged to keep passwords confidential and not allow anyone on their assigned terminal without logging off the system first. The classes were also informed that the Human Resources Section would maintain a master security table to monitor the use of passwords. Should any operator or backup want to change their password or terminal, they were to contact Human Resources.

## Implementation

Five days following the training, the new mainframe evaluation system went on-line. From the formation of the team to the on-line operation, this represented just over two months.

The current on-line system is designed to error check for invalid evaluation dates, wrong badge numbers for employees, raters or reviewers, incorrect employee location code, ratings without weights and weights without ratings. The system has a feature the previous system did not have, that allowing the operator to put an evaluation into an error file. It will be held there until corrected and then printed. Currently, all printing of the Employee Appraisal Report is done in triplicate by Data Processing Central Operations. On-site printing throughout the state on operator's printers is scheduled to be implemented within the next year.

The reporting capabilities of the on-line system will include individual rater modelling, predictive criterion inquiry, rater trend analysis, as well as rater comparison by unit, section, division, bureau and department. Other reports can be printed by rater, rank, classification, evaluation score or employee location.

## SOLUTION

Decentralize the existing automated performance evaluation system to a distributive processing mainframe application.

Such a solution represents yet another practical solution for more proficiently assessing the "Workforce 2000".

"Power to the Users: Decentralizing an Automated Evaluation System and Not Regretting It"

This panel focuses on how the Arizona Department of Public Safety converted a centralized automated performance evaluation system from an IBM-AT PC to an IBM 3081 mainframe system with distributive processing. Although this solution appears to be "bucking" the trends in automation, it has worked successfully for DPS. Such a solution has ensured greater flexibility, reduced turnaround, provided more accurate monitoring/tracking and developed an internal on-going system documentation and maintenance program provided by the Data Processing Division. The panel will discuss the conversion process and some of the new system's report generating features including predictive criterion inquiry, rater modelling, rater trend analysis and interdepartmental rater comparisons. Such practical solutions, as this one, will help Human Resources and Data Processing professionals more proficiently assess the "Workforce 2000".

ASSESSING THE IMPACT OF OFFICE AUTOMATION TECHNOLOGY  
ON SECRETARIAL AND CLERICAL JOBS

1990 IPMAAC Conference Presentation

by Marianne Bays

Organizational Consultant

Upper Montclair, NJ

INTRODUCTION

The implementation of office automation technology always presents both challenges and opportunities. Beyond the technical challenges involved with selection and installation of the most appropriate office automation hardware and software, there are management challenges that must be addressed if the technology is to fulfill its potential for improving organizational effectiveness. Among these are the technical training of employees, managing employee attitudes and expectations with regard to the technology and designing organizational structures to support both the initial implementation and the inevitable expansion of use of office technology over time.

In addition to the opportunity that office automation provides for increased organizational productivity (up to a tenfold return on investment according to one research study), opportunities for job enhancement can also result. Implementation of new technology will necessarily change job requirements, even when job responsibilities remain largely unchanged. If appropriate employee training and support is provided, the majority of employees can be expected to view the mastery of needed office automation knowledge, skills and abilities as a positive job challenge. If, however, inadequate training and support is provided, employees can become frustrated and the realization of benefits from the technology may be impeded. Research suggests that organizations that implement personal computer based office automation systems should expect the need to spend at least as much on training and support as they spend on the computers, themselves.

Finally, the need to design organizational structures to support both initial and expanded use of office technology can provide management with opportunities to enhance existing jobs and to structure new career options for employees. An increased need often exists for establishment of office technology specialist positions to support employee learning and resolve technical problems encountered. An increased need for positions focused on the day to day use of the technology in program activities can also arise. These needs are best met in some cases by establishment of new full time jobs in office automation career paths. In other cases, the organizations's needs may be met by restructuring existing jobs to include some new office automation support workload, through collateral duty assignments, or through establishment of task forces or standing committees. Often times, some combination of these approaches is most appropriate.



The focus of this presentation will be on the methodology used in and the results of a study completed in Region II of the U.S. Environmental Protection Agency during the spring and summer of 1989. This study examined the impact of current and planned implementation of office automation technology on the agency's 201 secretarial and clerical jobs.

#### ORGANIZATION'S TECHNOLOGY IMPLEMENTATION STATUS AT TIME OF STUDY

EPA Region II's office automation technology plan calls for the implementation of personal computer local area networks, with a 1-to-1 ratio of computers to employees. The ratio of computers to employees in the region at the time the study began was roughly 1-to-3, with about one third of secretarial and clerical employees having access to a personal computer at their workstation. Total professional information technology support staff in the region at that point included one manager and four specialists in the Information Services Branch. In addition, each division in the region had one employee (typically at a professional level, e.g., an environmental specialist or engineer) who was assigned collateral responsibility for office automation technology support in that division.

Standard software packages being implemented for use in the region included the WordPerfect word processing system; dBASE III, a data base management system; and LOTUS 1-2-3, a spreadsheet application. The office automation training curriculum in place at the time of the study included six Pace University courses designed and delivered for EPA employees:

Training Course	Course Length	Estimated % of Employees Trained	
		All jobs	Secr/Cler jobs
Introduction to Personal Computers	1 day	27%	46%+
Fundamentals of WordPerfect	2 days	19%	75%
Fundamentals of dBASE III+	2 days	9%	8%
Fundamentals of LOTUS 1-2-3	2 days	8%	3%
Intermediate dBASE III+	2 days	3%	-
Intermediate LOTUS 1-2-3	2 days	3%	-

#### MANAGEMENT ISSUES ADDRESSED IN STUDY

Implementation of new office automation technology stimulated a number of questions for EPA regional management. They wondered:

- 1) Would the implementation of the technology have grade impact on secretarial and clerical jobs?

- 2) What office automation proficiency levels were reasonable to expect of clerical and secretarial employees at different grade levels?
- 3) How much secretarial and clerical employee resistance to learning the new technology was operating in the region?
- 4) Was the existing training curriculum adequate to develop requisite levels of employee skills in use of office automation technology?
- 5) How could they assure adequate implementation support for the new technology given limited resources?

In addition, they wondered if the introduction of office automation technology might provide opportunity for job restructuring that would yield enhanced career opportunity for secretarial/clerical employees.

Management of the agency was particularly hopeful that there might be some way in which the technology implementation could result in job enhancement. Hindered by a noncompetitive Federal salary structure, especially in the highly competitive market of New York, the agency was experiencing difficulty attracting and retaining qualified clerical and secretarial employees.

#### METHODOLOGY

The study began with collection and review of written materials on secretarial and clerical jobs, technology implementation status and plans, and on the existing office automation training curriculum. A series of interviews were then held with EPA employees and management aimed at collecting information on office automation implementation issues and concerns. Following these interviews, a multi-purpose job analysis was conducted to provide the basis for identifying revisions needed in job descriptions, classification and selection/promotion procedures, and definition of curricula to meet new secretarial/clerical training needs related to the technology. Instead of a full job analysis, only office automation tasks and knowledge, skills and abilities were addressed. The focus in this analysis was on five job titles (i.e., GS-203 Personnel Assistant, GS-303 General Clerical and Administrative, GS-318 Secretary, GS-322 Clerk Typist, and GS-344 Management Specialist) at nine grade levels (i.e., GS-1/9).

Standard job analysis procedures posed some problems in this study. Use of the common practice of relying on position incumbents to serve as "subject matter experts" would have assumed that secretarial and clerical incumbents had the information needed about office automation task activity and proficiency requirements. In this case, because the information technology had not yet been fully introduced, this was not a reasonable assumption. Incumbents could tell us about their current job use of the technology and what it required, but could not tell us all that we needed to know about technology yet to be implemented.

In cases of new jobs or new job requirements, another fairly standard job analysis practice is to turn to the position supervisors for information. Again, this was not a fully satisfactory approach. Few supervisors had a clear concept of how they expected the technology to be implemented and to impact on the jobs of secretaries and clerks in their divisions. Many others were actually even more technologically inexperienced than the incumbents. Another approach was clearly needed.

The solution in this study was to work with a group in the region dubbed "office automation subject matter experts" to supplement data collected from incumbents and supervisors. The subject matter expert group assembled consisted primarily of the "PC Coordinators" who had collateral responsibility for office automation technology support in their divisions. While these individuals were not actually office automation "experts", they were more knowledgeable about use of the technology than were other employees. In particular, these individuals were the most knowledgeable about current and future office automation skill and knowledge requirements in the region. They were also able to provide the best information about office automation and information systems support workload that might be appropriately reassigned to secretarial/clerical positions, to provide enhanced job opportunities to employees.

Data collected on office automation tasks and their related knowledge, skills and abilities were compiled into a survey format and distributed to all 201 secretarial and clerical employees in the region. An employee attitude questionnaire was incorporated to assess employee feelings towards implementation of the new office automation technology. In addition, a set of background questions were added to the survey that enabled analysis of differences among employee subgroups with regard to training and classification issues. Usable responses were received from 80 employees (a 40% return).

## STUDY RESULTS

Key study results were as follows:

### KEY RESULT 1.

Generally, there is no grade impact on secretarial and clerical jobs resulting from implementation of new office automation technology.

The primary use of office automation technology by secretaries and clerical employees in the agency was found to be in support of performance of document preparation, document storage and retrieval and administrative reporting. These are standard secretarial and clerical work responsibilities that were already reflected in job descriptions and considered in job classification at EPA.

The conclusion drawn was that having office automation proficiency alone is no more grade relevant than having typing proficiency in secretarial and clerical jobs. To a point, these skills are required in order to accomplish tasks assigned. But, just as clerical and secretarial positions are not graded purely on the basis of typing proficiency, neither should the possession of office automation skills be viewed as grade controlling.

#### KEY RESULT 2.

Only broad grade distinctions needed to be made in establishing office automation proficiency requirements for existing clerical and secretarial jobs in the agency. Office automation task performance and related proficiency levels were found not to vary significantly by job title or specific grade.

Office automation tasks and their associated knowledge, skill and ability requirements were tentatively categorized as "Basic", "Intermediate" or "Advanced" based upon the current frequency of task performance reported by employees. These categorizations were then adjusted and finalized based upon the input of the "Office Automation Subject Matter Expert Group" and consideration of which office automation knowledge, skills and abilities were "required" (i.e., needed to perform common required job tasks) vs. "desirable" (i.e., not required, but expected to enhance task performance if possessed).

Despite the finding that there were little or no current grade differences in office automation task performance nor in office automation proficiency levels, other considerations led to the decision that office automation proficiency requirements should be established for existing jobs by three grade "bands" (i.e., GS-1/3, GS-4/5 and GS-6 and above.

The key consideration here was the recognition that office automation use was so new in the agency that the proficiency distinctions between lower and higher grades that one would reasonably expect to develop over time had not yet had time to occur. Therefore, thought had to be given to how office automation proficiency could be expected to grow over time and how this would relate to secretarial and clerical career paths.

Our analysis led to the following conclusions. Region II hires most of its clerical employees at the GS-1/2/3 levels. At these grades (especially given current recruiting difficulties), no office automation proficiency could reasonably be expected at the time of hire. Once hired, however, the majority of these employees are expected to gain the skill requirements to progress noncompetitively to GS-4. The transition from GS-3 to GS-4 was, therefore, established as a key point in the career path for the the purpose of defining office automation proficiency requirements for clerical jobs in the region. Similarly, advancement from GS-5 to GS-6 was identified as a career transition point. Promotion opportunities from GS-5 to GS-6 are fairly limited and only the most skilled clerical and secretarial employees move to this level.



These considerations led to the recommendation that the office automation proficiency requirements be established for existing jobs by grade "bands" as follows:

Basic Office Automation Proficiency - This was set as the level of proficiency that all secretarial/clerical employees would be expected to possess in order to effectively perform the most common office automation tasks. At GS-1/2/3 levels, job incumbents would need to be able to apply these KSA's with assistance. For promotion to GS-4 and at the higher grade levels, job incumbents would be expected to be able to apply these KSA's independently.

Intermediate Office Automation Proficiency - This was set as the level of proficiency that all secretarial/clerical employees at GS-4 and above would be expected to possess in order to more effectively perform the most common office automation tasks in their jobs. At GS-4/5 levels, job incumbents would need to be able to apply these knowledge, skills and abilities (KSA's) with assistance. For promotion to GS-6 and at the higher grade levels, job incumbents would be expected to be able to apply these KSA's independently.

Advanced Office Automation Proficiency Requirements - The knowledge, skill and ability requirements for office automation tasks that were not currently performed with great frequency formed the basis for this proficiency category. These infrequently performed tasks were viewed as providing the foundation for restructuring existing jobs or establishing new jobs at GS-6 or above to focus on office automation implementation and information system use in the region. The specific subset of advanced KSA's required in an office automation support job would depend on the advanced office automation tasks specifically assigned. In addition, general office automation proficiency requirements established for all support jobs at GS-6 and above included expert levels of the Basic Office Automation Proficiencies and independent ability to use the Intermediate KSA's.

### KEY RESULT 3.

Findings were that the overwhelming majority of secretarial and clerical employees view implementation of new office automation technology as a positive opportunity both careerwise and for the organization.

Regardless of grade, work location or title, secretarial and clerical employee attitudes and feelings towards office automation were found to be generally very positive. Employees reported that they believed that office automation WILL improve their work capabilities, increase job challenge, improve productivity and quality within the region, be enjoyable to work with, improve communications within the region, help service the public, and provide them with increased job opportunities both inside and outside of EPA.

They also generally reported that they believe that office automation WILL NOT be stressful, have harmful health effects, prove to be a mistake, be embarrassing, impede productivity or be difficult to master. This suggests that as long as employees are provided with the necessary training, equipment and "practice time", the region can expect that they will, for the most part, be self-motivated to develop the required levels of proficiency in new technology.

#### KEY RESULT 4.

While the agency's available office automation training curriculum was found to be providing a sound basis for acquisition of many of the most commonly needed office automation proficiency requirements, a need for new or revised training at GS-4 and above was found. A need was also identified for supervisors to provide employees with ready access to a PC after training so that they can practice and hone skills taught.

The secretarial and clerical employees who responded to the Office Automation Task and Skill survey provided data on what their current level of mastery of 76 different office automation knowledge, skills and abilities. They also reported where they had obtained their level of proficiency. This allowed us both to determine where particular KSA deficiencies might exist and whether existing training curriculum could be relied upon to increase proficiency levels to required levels.

One additional piece of information collected from employees proved especially valuable in this analysis. Survey respondents were all asked to report whether or not they had a PC at their work station reserved primarily for their use. This allowed analysis of the impact of ready access to a PC upon which to practice skills taught in training.

A multiple step analysis was conducted to determine the level of office automation proficiency assured by course attendance.

- In the 1st step, average level of possession of KSA's for all employees was calculated.
- In the 2nd step, respondents were broken into groups of those who had attended relevant courses and those who had not and the average level of possession of KSA's were calculated for each subgroup.
- In the 3rd step, an employee's access to a Personal Computer was added to the analysis. The groups who had attended relevant training were further divided into two, those with a PC at their workstation reserved for their use and those without. Average level of possession of KSA's was calculated for each of these subgroups.
- In the 4th and final step, the KSA's that at least 50% or more of respondents who had attended a particular training course attributed to that course were identified.



The highest levels of basic office automation proficiencies were most often related to training attendance coupled with ready access to a PC upon which to practice. While this was not a surprising finding in light of what we as personnel professionals know about good training practice, this finding was probably the one with the greatest immediate practical value to EPA Region II. While only about 30% of survey respondents had access to a PC at their work station at the time of this study, at least 75% of them had already attended some form of office automation training. At a cost estimated by the agency at \$100 per training day (and not including employee salary costs for time spent in training), this minimally suggests that about \$9,000 of training expense, if not wasted has, at least, been ineffectively spent.

A general weakness in proficiency levels with regard to knowledge of different printer capabilities, Local Area Network (LAN) functionality, communications software and hardware configuration, and electronic mail system (EMAIL) use were noted for employees at all grades.

In addition, intermediate KSA training needs not met by the existing training curriculum included:

- Knowledge of PC care and security policy and skill in use of basic PC/DOS commands and procedures
- Skill in using "intermediate" functions of WordPerfect (e.g., document conversion, line drawing, printing mailing labels, printing documents horizontally, etc.)
- Knowledge of PC database recordkeeping applications, and skill in report retrieval and record updating from such systems

Final office automation training revision and development recommendations included:

- Development of two new formal training courses (PC and PC/DOS Basics and Intermediate WordPerfect).
- Minor revision of the existing WordPerfect training.
- Establishment of systematic on the job training and/or informal briefings in certain areas (e.g., LAN, EMAIL, printer capabilities, etc.)
- Targeted work assignments for GS-4/5 level employees to help insure that required levels of proficiency are developed for advancement to GS-6.
- Refresher training in WordPerfect for employees who attended training prematurely, before they actually had access to a Personal Computer

#### KEY RESULT 5.

Study results clearly showed that the implementation of office automation technology in this agency can provide opportunities for job and career enhancement through the assignment of new workload to jobs. Each division of EPA Region II now has or will have a need for additional office automation support, although the specific type and degree of this varies. The workload of concern will be greatest in the larger divisions, especially where the designated PC Coordinator is already unable to devote sufficient time to accomplishment of his or her collateral assignment.

The study generated a list of specific job activities involving office automation technology implementation and information system use that could be possible task components of new job structures. This was based upon the aspect of management and "office automation subject matter expert" interviews that focused on the question, "What work is currently not being performed or is being performed by professionals that could be used to create new office automation support positions?"

The specific type of office automation support work needed depends both upon the size and the particular work of each organization. Some offices' greatest need is for increased administrative support in the implementation and use of administrative office automation systems. Others have a greater need for increased attention to the adaptation and use of information systems to support specific technical program activities. Still others have need for increased support in both areas. For this reason, a support structure that provides management with the flexibility to act on the particular office needs was recommended.

Several different options for assignment and staffing of new office automation work functions were provided to management. These included:

- An on-the-job training approach wherein a relatively simple and light office automation workload would be assigned to existing employees on a rotating basis in order to both get the necessary work done and to provide equal opportunity to employees to gain new proficiencies.
- A restructuring of existing jobs to include assignment of new job functions on a more permanent basis to qualified employees in cases where the office automation support workload is more complex and greater continuity in task performance is needed.
- Creating and competitively filling new full time office automation support jobs with advancement potential in cases where the new support workload requires full time attention and is expected to increase in complexity over time.
- A combination of more than one of the above approaches.

Guidelines for establishing office automation career path jobs were included in the final report study. Two possible job titles were suggested: Office Technology Assistant GS-303-6/7 and Environmental Protection Assistant GS-029-6/7. The GS-303 series was felt to be appropriate for jobs involving a variety of administrative duties in support of office automation technology implementation and use. The GS-029 series was felt to be more appropriate for those jobs involving performance of a variety of duties in support of EPA technical program activities, where most involve use of information systems.

Differences in career development opportunities provided by these two types of jobs were noted. The jobs classified in the GS-029 series could provide a "bridge" to more technical positions in the agency because specific technical knowledge should be gained in these jobs. Those jobs classified in the GS-303 series should, on the other hand, provide higher level administrative skills to incumbents and can be viewed as specialized extensions of the existing career path for secretarial and clerical employees in the agency. Both types of jobs might also provide career development opportunities leading to qualification for technical assignments in the region's information systems branch, as well.

The final recommendation in this part of the study was that management engage in analysis of their specific needs for additional automation support, consider available options, and then formulate specific action plans for assignment and staffing of new office automation work functions within each division.

#### BEYOND THE STUDY

A follow up interview was held with EPA management, 6 months after study completion. Status of implementation of study recommendations at that point was found to be as follows:

- The process of supervisory review of office automation proficiency levels prior to noncompetitive promotion actions for clerical/secretarial job incumbents has been successfully implemented.
- A few existing jobs have been restructured to add additional office automation support workload and, in some cases, this has resulted in position upgrades.
- The first new office automation support career path position is now being established, using the Office Technology Assistant GS-303 title.
- Training has been revamped much as recommended. Completion of the agency's PC lab and training facility has facilitated their ability to offer additional office automation training.

- The study effectively drew attention to the inefficiency of sending employees to training without providing access to a personal computer on which to practice skills taught. However, no formal monitoring of this has been implemented.
- While an attempt was made to coordinate action planning across the region for the assignment and staffing of new office automation functions, Senior Management did not follow up on this. Implementation has, therefore, been individual to branches and not very systematic.

#### REFERENCES

Special Report - Office Automation: Making it Pay Off.  
Business Week, October 12, 1987.

Managing Personal Computers in the Large Organization. Report prepared for Lotus Development Corporation by Nolan, Norton & Company, 1987.

PERFORMANCE ON A PC-BASED TYPING TEST (R.D. CRAIG)  
VERSUS PERFORMANCE ON A TRADITIONAL TYPING TEST

BARBARA E. LEIGHTON  
JADE KUAN HOFFMAN  
THUNG RUNG LIN  
Personnel Selection Branch  
Personnel Commission  
Los Angeles Unified School District

Performance on a computer-based typing test (R.D. Craig) was compared to performance on a traditional typing test (administered on an electronic typewriter). The study examined the extent that previous typing experience and specific machine experience would influence performance on each machine. Test retest reliability figures were established on each machine. Equivalent scores were established for tests on each machine. Other factors related to participants' perceptions of ease, clarity of instructions, and general comments were also analyzed.

In recent history, technological advances have dramatically changed equipment available for producing typed copy. Typists may use a variety of typing instruments, including computers and electronic typewriters. Features available on machines can affect the traditional methods used for typing. Typists who have become accustomed to typing on a computer can make instant, non-detectable corrections for errors. They may be able to produce high quality final products in a timely manner when using a word processing package, but be unable to pass a traditional typing test when no corrections are allowed.

Technology has also influenced the selection processes used to assess typing skills. In particular, PC programs for assessing typing skills are available from numerous vendors. PC-based skills assessment programs offer potential advantages such as greater standardization, better control of timing, and automated counts for character production and for errors. A typing test administered on a Personal Computer also increases face validity for jobs which involve computer usage.

While significant advantages do exist, potential problems associated with PC-based typing tests should not be ignored. Typing tests may use different approaches for calculating error and text production counts, so it may be difficult to equate scores across methods. In particular, this problem can arise when there is an existing words-per-minute standard for scoring typing speed and accuracy in a traditional test.

The study was initiated while testing for an Office Computer Operator classification. In this case, candidates are required to demonstrate a 30 wpm net typing speed in order to qualify for the classification. Due to limited typing instruments available for testing purposes, candidates in previous administrations were tested on electric or electronic typewriters.

A wide range of typing tests and formats have been used to assess job related typing skills. Administration methods, scoring formulas, treatment of errors, source material, formats of tests, time allowed, and equipment used can all affect comparisons of typing test scores. Test formats may involve straight text typing, numeric typing or an alpha/numeric combination. Tests may focus on specific skills in tabulation, proof reading, completion of forms, or layout of the final documents. Many typing tests use typed copy as source material.



however, handwritten copy or audiotapes may also be used (West, 1969; MAPAC, 1990). Tests can involve a series of projects that must be completed in a given period of time, or time required for completion can be one rating factor. A MAPAC survey of clerical testing practices received responses from 76 agencies and companies throughout the nation. Results indicated that 53 respondents test typing skills by requiring typists to copy typed source material, 52 respondents use a straight copy format, and while 11 use PC terminals for testing, 57 use electric typewriters (MAPAC, 1990). Straight copy formats allow speed and accuracy to be measured quantitatively with relative efficiency.

West (1969) has compiled extensive material on typing techniques used for teaching typing skills and for testing. He strongly recommends giving separate scores for speed and for accuracy skills. He states that "there has been a uniform finding ... of an essentially zero relationship between stroking speed and stroking accuracy. Typists at all levels of speed are found at all levels of accuracy" (West, 1969 p.238). Using this strategy, a final composite score can be given while requiring a minimum score for speed and allowing a maximum number of errors for each level of speed. A scoring chart can be used to show the maximum number of errors allowed for each level of gross words per minute speed. For example, a margin reading might show that 40 words per minute had been typed in 17 lines of typing. This would be a passing score if 8 or fewer errors had been made.

Another common practice in the treatment of errors is to multiply total errors, or errors per minute, by a pre-determined factor. This error figure is subtracted from the speed measure of gross words per minute in order to derive a final score. The error factor used generally depends on the cost of time required to fix the error. The computerized typing test used in this study recommended an error factor of 3 for each error per minute. This factor was based on the assumption that an average error could be corrected in the amount of time that it would take a typist to type 3 words (K.B. Craig, 1989). Other traditional typing tests, such as the traditional test in this study, multiply errors per minute by 10. Errors made on machines that do not have correcting capabilities would cost substantially more time to correct. A more moderate error factor of 5 is used by other agencies (Arco, 1987).

Previous studies have found that skilled typists detect about half of their errors immediately (Cooper, 1983). Most traditional typing tests do not allow corrections to be made during the testing process. However, technological changes in typing instruments have brought about computers and various electronic typewriters that can produce perfect form final products even when errors were corrected during work. The computerized typing test used in this study assumes that equipment used on the job will allow non-detectable error corrections. In this test, immediate backspace corrections are allowed.

The use of computers for testing computer-based typing positions has strong face validity. Computerized testing of other typing positions may raise concerns. Specific studies that compare computer typing tests to traditional typing tests were not found in the literature of this field, however, a number of studies have compared computerized tests to paper and pencil tests. Olsen, Maynes, Slawson and Ho (1989) compared test results for three different testing formats: a paper and pencil test, a computer administered test, and a computerized adaptive test. These formats were administered to 350 third grade students and 225 sixth grade students. No significant differences were found in test results in this study.

Other studies have examined the issue of computer anxiety that some people report when required to use a computer in a testing situation. Ward, Hooper and Kinnafin (1989) randomly assigned 50 students to take either a computerized test or a paper and pencil test. While there was no significant difference in test



performance, a significantly higher level of anxiety was reported by students who were assigned to the computer test. In this case seventy-five percent felt the computer test was harder than the paper and pencil test.

Herkimer (1985) studied how typing ability, among other factors, would correlate with computer anxiety. The study was conducted in a naturalistic setting, using subjects from a business environment. In this study, typing ability correlated positively with computer anxiety only for older males with limited typing ability. This suggests that computer anxiety need not be considered a significant factor when using computers to assess typing skills.

Computerized typing tests have been adopted by a number of companies and agencies; others are in the process of developing or purchasing computerized typing tests. The R.D. Craig brochure lists over 90 companies and agencies that have used the Speed and Accuracy Typing Test described in this study. Some selection planners implement computerized tests in ready-made form, others make revisions to existing software packages, and some have special software packages designed for their unique testing program.

This study was conducted to compare a ready-made package to the traditional typing test used for selection in a large school district in southern California. Due to restrictions in resources and established classification requirements, the likelihood of a complete conversion to the computer typing test program is low. However, potential does exist to implement the computer typing test in addition to the traditional typing test. Therefore, the determination of equivalent scores on the two test types was a primary focus in this study.

#### DESIGN

The two tests compared in this study are both speed and accuracy typing tests which require participants to copy from typewritten text. The only major adjustment that was made to the computer based test before the study was to lengthen the practice time from 2 minutes to 7 minutes. This was done to maintain consistency with practice time allowed for the traditional typing test. Table 1 summarizes some similarities and differences between these two tests.

The four main differences between the two tests were:

1. immediate corrections were allowed on the computer test but not on the traditional test
2. the computer test had wrap-around margins so that participants did not need to press the return/enter key; the traditional test required exact line for line typing
3. the error factors used in scoring were different
4. text used was different for each machine.

Participants were asked to fill out a questionnaire during a break period toward the end of each study administration. Please refer to Appendix A for a summary of information gained from the questionnaire. This instrument was used to obtain information relating to the extent of typing background and machine experience of participants.

Based on the background review of literature, the R.D. Craig brochure and questions raised by the researchers, the following hypotheses were tested during the study.

1. H1 When immediate corrections are allowed, typists will make about half as many errors on the computer test as they make on the traditional test.
2. H2 Test re-test reliabilities will be the higher for the computer-based test.
3. H3 There will be significant differences in scores between groups according to their previous experience with typewriters and computers.

TABLE 1

COMPARISONS BETWEEN TRADITIONAL VS. CRAIG PC-BASED TYPING TEST

(Factor)	Traditional	(Craig) PC-Based
7-minute practice 5-minute test	Yes	Yes
Gross wpm= characters typed/5	Yes	Yes
Errors/minute (EPM)= total errors/5	Yes	Yes
Error term	EPM x 10	EPM x 5
Net words/minute= gross wpm - error term	Yes	Yes
Administration	Oral instructions given by proctor	Instructions on screen reviewed orally by proctor
Scoring	Hand scored	Computer scored immediately after test
Timer	Test timed by proctor with a wind-up timer/bell	Computer started time when typist started typing - shut off automatically
Text	Type line for line. Must hit return key at end of each line No warning bell for line end. Tab for paragraphs.	Enter key (return) non-functional for test. Automatic wrap-around for line ends. No tabs or paragraphs. Lines ends of text did not match line ends of screen image.
Average word length	5.8	6.0
Corrections	No corrections.	Corrections allowed by immediate backspace only.
Accuracy measure	Part of net score	Expressed as a separate percentage = net wpm/ gross wpm x 100.
Equipment used	Panasonic Electronic typewriters - no correcting tape.	IBM PCs with enhanced keyboards

In each data-collecting session, half the participants were randomly assigned to take the computer test first. The remaining half took the traditional test first. This test order was used as the treatment effect for the study design. Specific information regarding study administration is covered in the following section. Data were analyzed using the Statistical Package for Social Sciences - PC Version (SPSS/PC) (Norusis, 1988).

## METHOD

### Subjects

The sample for this study consisted of volunteer participants who became involved as a result of several recruitment strategies intended to reduce the restriction of range. In all cases, the emphasis was to recruit a variety of people with different typing abilities and backgrounds. Some participants became involved as a result of bulletins and specific contacts through School District offices located in the vicinity of the Personnel Selection Branch. Other participants responded to an advertisement placed in a district-wide newsletter. Targeted recruitment was also used to encourage the participation of Office Computer Operators, Word Processing Operators, and students who work with electronic typewriters in a typing class. A total of 178 people generated usable results for this study.

Two participants began the study exercises but dropped out before completing them. One had been "volunteered" by a supervisor and refused to actively participate in the typing exercises. The other was interested in one trial on the computer typing test but did not follow through with the other exercises. All participants were told that their scores would be confidential and would not adversely affect their current job.

Up to 8 participants were scheduled for each study session. An entire session required no more than one and one half hours to administer. The study was conducted in the same room that is used to administer actual traditional typing tests for the school district. Electronic Panasonic typewriters used for testing were used as the typewriters in the study; IBM PCs were used to administer the computerized typing test.

Participants were able to choose chairs of appropriate height as they entered the room and found typing stations. Once participants were seated at typing stations, a brief introduction was given. Typing materials were distributed and the instructions for the traditional typing test were announced to all participants. Then a coin was tossed as a means of randomly assigning half the group to computer stations. On screen instructions for sign-on and the computer test instructions were reviewed orally as people at the computer stations followed along. Participants were encouraged to ask questions if they didn't understand instructions or had difficulty with functions on either machine.

All participants beginning at the computer were instructed to start typing after they had read through the instructions and felt ready. The timer for the traditional typing test was not set until all computer testers had started their automatically-timed administration.

Participants were asked to switch machines after the practice test and actual test on the first machine (i.e., people who had been tested on the computer initially moved to their typewriter stations). This alternating between machines was done to avoid an extended practice effect on individual machines.

On the second machine, participants also took a practice test and actual test. After a brief break, participants returned to their original machines for a retest. The questionnaire was completed during a second break period before the final retest. Appendix A indicates the times allowed for each test part.

The traditional typing tests were scored by hand. In contrast to many selection typing tests, the scores were calculated even when negative figures resulted. Scores plotted into a positive correlation that will be described below.

## RESULTS

As expected, there were significantly fewer errors made on the computer test. A dependent t-test of accuracy expressed as a percentage (net wpm score/gross wpm x 100) indicated that the mean for the traditional typing test was 41.8% accuracy rate while the computer test accuracy rate was 91.2% ( $t(174) = -15.77, p < .01$ ).

Lowest total error scores for each of the two computer trials and two traditional trials were compared using a dependent t-test. The mean total errors on the traditional test, with no corrections allowed, was 11.96, while the mean total errors on the computer test, with corrections was 6.07 ( $t(175) = 10.27, p < .01$ ). This substantiates Cooper's statement that typists tend to detect about half their errors immediately. (Cooper, 1983)

The immediate test-retest reliability of the traditional instrument as  $r = .74, p < .01$ . The t-test result,  $t(157) = 1.12, p = .27$  two-tailed, indicated that test performance did not differ significantly between the first and second trials. However, the original 21 cases did not have the traditional test order recorded, therefore, those scores were not included in this calculation.

The immediate test-retest reliability for the PC-based instrument was  $r = 0.95, p < .01$ . The t-test result indicated a significant difference between the first and second trials. The mean difference was  $-.93$  with a slight improvement on the second trial  $t(173) = -2.43, p = .02$ , two tailed.

On the basis of the questionnaire responses, two groups were established according to typing equipment used most. There were 52 respondents who reported using only typewriters most, while 88 reported using only computers most. The scores for net words per minute, gross words per minute, and accuracy on each machine were compared using t-tests for these groups. An adjusted net words per minute score for each machine was calculated by using the average error factor between the two machines. The computer scoring system used an error factor of 3 x errors per minute (epm) while the typewriter scoring system used a factor of 10 x epm. In order to equate error factors, these two factors were averaged and adjusted scores were computed using the average error factor of 6.5 x epm.

Table 2 lists results of these t-tests. In this case, a machine effect was found for people who reported using computers most often. On the computer test, this group had significantly higher scores for accuracy and net words per minute, even when the average error factor was used. The differences between gross words per minute were not significant. Also there were no significant differences between groups for any of the scores from the traditional typing test on the electronic typewriter.

In order to determine equivalent test scores, the traditional net wpm score for all participants was used as an independent variable (X) and PC net wpm score was used as the dependant variable (Y). Regression analysis produced the following results:

Constant	30.3215
R Squared	.5799
No. of Observation	177
Degrees of Freedom	175
X Coefficient	.7615
Standard error	10.4634

Table 2  
T-TEST OF SCORES BY MACHINE EXPERIENCE

Dependent variable	Machine Used Most:				t	p
	Computer		Typewriter			
	mean	std	mean	std		
Traditional Gross wpm	49.6	15.4	46.1	13.0	-1.38	.17
Traditional Accuracy %	45.7	35.4	35.1	60.3	-1.31	.19
Traditional Net wpm	23.6	17.4	18.9	24.2	-1.32	.19
Adjusted Traditional Net wpm	32.1	15.6	28.0	19.0	-1.38	.17
PC Gross wpm	50.9	15.6	46.0	13.5	-1.89	.06
PC Accuracy %	92.7	9.5	88.3	11.8	-2.46	.02*
PC Net wpm	47.1	15.7	40.5	15.1	-2.43	.02*
Adjusted PC Net	43.1	17.0	34.9	17.7	-2.71	.01*

Note: Computer n = 88  
Typewriter n = 52

All scores are the best out of two trials. \* p < .05

Regression Equation:  $Y = X \times .7615 + 30.3215$

The regression equation can be used to estimate the PC-based net score one could expect from a typist if the traditional net score is known. The equation can also be used to estimate equivalent norms for the PC and traditional tests. Table 3 is a norm table which projects PC scores from traditional scores using this equation.

Regression Equations: for 3 x EPM:  $Y = X \times .7615 + 30.3215$   
for 5 x EPM:  $Y = X \times .7540 + 26.5550$   
for 10 x EPM:  $Y = X \times .5716 + 13.9441$   
Degrees of Freedom: 175

Computer and typewriter scores were also compared for gross words per minute (gross wpm), accuracy, and net words per minute (net wpm).

The gross words per minute score was used to measure speed for both machines. It was expected that affects on this variable would be minimal because this score is calculated before error calculations are made. However, a mean difference of .88 was significant at the p < .05 level with a dependent t-test. Using the best speed score of the two trials, the mean of gross words per minute on the traditional typewriter was 47.84 while the computer mean revealed slightly faster typing (m = 48.72).

Table 3

## NORM TABLE OF EQUIVALENT SCORES

Predicted net computer scores based on net traditional score.

Traditional Percentile	Traditional Net Score	Net Computer Scores Predicted Using Different Error Terms		
	(10 x EPM)	(3 x EPM)	(5 x EPM)	(10 x EPM)
13	6	30.3	26.6	13.9
21	5	34.1	30.3	16.8
27	10	37.9	34.1	19.7
31	15	41.7	37.9	22.5
43	20	45.6	41.6	25.4
50	25	49.4	45.4	28.3
60	30	53.2	49.2	31.1
69	35	57.0	53.0	34.0
79	40	60.8	56.7	36.8
89	45	64.6	60.5	39.7
94	50	68.4	64.3	42.5
100	55	72.2	68.0	45.4

Before changing error factors, Pearson's correlation coefficient between Traditional and PC net words per minute scores was  $r=0.75$ ,  $p<.01$ . Even though the correlation seems high, a dependent  $t(176) = -22.00$ ,  $p<.01$  two-tailed, indicated that test performance differed significantly. As a group, participants received higher typing scores using the PC-based instrument.

After deleting outliers and converting scores to the average error factor, a dependent  $t$ -test, indicated that scores still differed significantly;  $t(175) = -12.23$ ;  $r=0.83$ ,  $p<.01$ .

One of the questions on the questionnaire asked, "Which machine was easier for you to use today?" Participants were given the opportunity to check off: a. electronic typewriter, b. computer, c. both. Only 2.5% indicated that the electronic typewriter was easier to use. An overwhelming 82.4% said that the computer was easier to use, 10.1% checked the category marked "both" and 5% did not respond.

An independent  $t$ -test indicated that there were no significant differences between net wpm scores for treatment groups. Scores were not affected by assignment to the computer or the typewriter test for the first test part. As previous studies have shown, there was no significant correlation between typing speed and typing accuracy.



## DISCUSSION

The sample for this study comprised a wide range of individuals with a variety of backgrounds. Because of the intention to develop a norm table of scores, people with all levels of typing ability were sought. More than half of the sample (54.7%) was made up of people from classifications which require typing projects or speed and accuracy typing skills of 30 net wpm or more. The sample of 178 people represented a cross section of the major ethnic groups with 12.6% Asian, 37.1% Black, 29.1% Hispanic, and 21.1% White.

In order to maintain study consistency, only Panasonic electronic typewriters were used to measure typing on the traditional test. Variation between models of electronic typewriters and electric or manual typewriters could be measured in the future. The norm table developed for this study applies to the electronic typewriters used for centralized testing in the school district.

Results from the questionnaire showed that 86.8 % of the participants felt the PC was easier to use. This study focused on perceived 'ease of use' rather than anxiety level. It was assumed that in an actual testing situation which allowed practice periods and/or two test trials, people would feel lower anxiety when using a machine perceived as easier to use.

Most people did not write additional comments. Of the 57 comments that were made, 25 involved negative comments about the typewriters; 2 negative comments were made about the computers. Positive comments about the computer test were made by 7 people. There were no positive comments specific to the typewriter test. The overall lack of complaints about the computer testing compared to typewriter testing serves as additional justification for using a computer test as an alternative to the traditional typing test.

Whether or not corrections can be made and how errors are actually treated have strong effects on final typing test scores. Today especially, appropriate selection for typing positions depends heavily on job analysis. As office equipment changes, the skills which enable an employee to work most effectively can also change. If typing equipment on-the-job allows immediate non-detectable corrections, then a typing test which allows corrections and uses a low error factor might be indicated. If typing errors are costly to correct, for example carbon paper forms are used for information that becomes a permanent record, then accuracy would be a more critical area of testing.

Some typing positions involve typing assignments that need to be completed in a given period of time. During this time, there are interruptions for other tasks such as answering the telephone, filing or retrieving information. The speed of a production typist may not be as critical as the ability to produce a final product in an acceptable period of time. In such a case, a typist who can self-edit may be more valuable than a fast typist, especially if the fast typist makes numerous errors. Computer tests can be adapted to meet the needs of a variety of testing situations.

Separate scores for speed and accuracy allow typing tests to be used in a number of ways. Separate cut-offs can be set to qualify typists for positions which require different typing skills. Composite scores can be used to equate scores between a test that allows corrections and one that does not. In addition, composite scoring can also be used to convert scores from tests that use different error factors.

These concerns are important to consider as typing positions are affected by technological advances in office machines. Selection specialists should be prepared to compare tests given on different typing instruments, compare test scores which result from different error factors and make assessments about when, if ever, corrections should be allowed on typing tests. By taking these measures selection specialists can revise tests for typing positions as needed.

47

This process of test evaluation will help to insure that qualified employees are not excluded from the selection process and that those selected meet the current needs of the job.

#### References

- Arco Editorial Board. (1985). Civil Service Typing Tests. New York: Arco Publishing, Inc.
- Cooper, William E. (1983). Cognitive Aspects of Skilled Typewriting. New York, Heidelberg, Berlin: Springer-Verlag.
- Craig, R.D. (1989). R.D. Craig Assessments Speed and Accuracy Typing Test Manual. Midland, Ontario: R.D. Craig Assessments Inc.
- Herkimer, Brenda. (1985). "Computer Anxiety as State Anxiety and Time-on-Task and Their Relationship to Sex, Age, Previous Experience, and Typing Ability." (Doctoral dissertation, University of Southern California).
- MAPAC Committee on Clerical Testing. (1990). "MAPAC-Sponsored National Survey of Clerical Testing Practices." MAPAC News, January, 1990.
- Norusis, M.J. (1988). SPSS/PC+ V2.0 Base Manual. Chicago, IL: SPSS Inc.
- Olsen, James B.; Maynes, Dennis D.; Slawson, Dean; Ho, Kevin. (1989). "Comparisons of Paper-Administered, Computer-Administered and Computerized Adaptive Achievement Tests." Journal of Educational Computing Research, 5, 311-326.
- Ward, Thomas J., Hooper, Simon R. and Hannafin, Kathleen M. (1989). "The Effect of Computerized Tests on the Performance and Attitudes of College Students." Journal of Educational Computing Research, 5, 327-333.
- West, Leonard J. (1969). Acquisition of Typewriting Skills, Methods and Research in Teaching Typewriting. Great Britain: Pitman Publishing Corporation.

#### Footnote

The authors gratefully acknowledge Calvin Hoffman for his helpful comments and assistance in reviewing this paper. However, full responsibility for the contents of this paper remains with the authors.

APPENDIX A

SCHEDULE FOR TYPING STUDY ADMINISTRATION

15 minutes      INTRODUCTION  
Ask people to find suitable chair at typewriter.  
Pass out paper for typewriters.  
Demonstrate paper loading technique.  
Give introductory information about study.  
Toss coin to split group.

7 minutes      PRACTICE PERIOD (Computers and typewriters)

5 minutes      ACTUAL TEST

SWITCH - People on computers for first part move to typewriters.  
people on typewriters for first part move to computers.

7 minutes      PRACTICE PERIOD (Computers and typewriters)

5 minutes      ACTUAL TEST

SWITCH - First people on computers move back to computers.  
First people on typewriters move back to typewriters.

ROSTER - sign in. (short break - 5 minutes)

5 minutes      ACTUAL TEST

SWITCH - All move back to second machine used.

QUESTIONNAIRE (short break - 5 minutes)

5 minutes      ACTUAL TEST

# APPENDIX B

## SUMMARY OF INFORMATION GAINED FROM STUDY QUESTIONNAIRE

	ELECTRONIC TYPING TEST	COMPUTER TYPING TEST	BOTH	VALID CASES
Instructions				(178)
CLEAR	98.9%	97.7%	-	
LESS THAN CLEAR	1.1%	2.3%	-	
Enough practice time				(166)
YES	88.6%	94.6%	-	
NO	12.0%	5.4%	-	
Machine easier to use.	2.4%	87.6%	10.0%	(170)

The median and mode for most recent use of a typewriter or computer keyboard, measured in weeks, was 1 week. (157)

Equipment used most: (173)

48% PERSONAL COMPUTER      5% MANUAL TYPEWRITER  
25% ELECTRIC/ELECTRONIC TYPEWRITER  
15% PC AND TYPEWRITER      5% OTHER

COLUMN A = used daily or for 4 or more hours in the last week

COLUMN B = used at least once in the last week

COLUMN C = used regularly in the past, but not in the last week  
note: regularly = daily, or at least 4 hours per week

COLUMN D = used at least once in the past, but not in the last week

	A	B	C	D	
IBM PC	37%	6%	7%	12%	(111)
OTHER COMPUTER	15%	7%	14%	17%	( 94)
ELECTRONIC TYPEWRITER	18%	13%	7%	15%	( 95)
ELECTRIC TYPEWRITER	14%	7%	10%	14%	( 78)
MANUAL TYPEWRITER	2%	2%	2%	26%	( 52)

Job Analysis Across Two Cultures -- U.S. and Japan:  
Collecting Accurate Data Without the Use of Job Incumbents

Kevin G. Love, Ph.D.  
Department of Management  
Central Michigan University  
Mt. Pleasant, MI 48859

Abstract

A job analysis was completed for the position of small motor assembler within a U.S.-Japan joint venture manufacturing organization. In response to a "typical" Japanese management climate and relevant legal and statistical personnel selection system requirements for the U.S., a "hybrid" job analysis procedure was developed to provide the necessary foundation for the development of a behavior-based assembler selection system.

The job analysis utilized features from several established job analysis methods to overcome obstacles to data collection which included: the absence of an existing organization and job incumbents, the desire of both U.S. and Japanese management to go beyond paper and pencil testing for selection purposes, and significant cultural differences which prohibited the use of traditional job analysis data collection (i.e., employee questionnaires).

Table 1

Requisite Performance Areas for Small Motor Assembler

**ATTENTION TO MAINTENANCE AND SAFETY**

Candidate displays a willingness to participate in maintaining equipment and a clean and efficient work station. Demonstrates an understanding of the importance of following safety procedures while working. Indicated by:

- returning tools to proper place
- placing unused parts into appropriate bins/areas
- thoroughly cleaning assigned area
- following safe procedures on various assembly and/or soldering operations

**TEAM ATTITUDE AND PARTICIPATION**

Positive interaction with work team members showing support and encouragement to accomplish team goals. Displays skills necessary to foster and maintain a positive cohesive work group. Indicated by:

- physically and psychologically assisting team members in their work
- showing sensitivity towards others' thoughts and feelings
- critiquing other team members' work in a positive manner

**PLANNING AND ORGANIZING WORK**

Studying required materials (e.g., schematic drawings, manuals, reports, etc.) before beginning work operation (assembly) to plan for work set up and facilitate the assembly procedure. Prepares for work by checking to see if machines, tools, etc. are in proper condition and working order and needed parts, materials, etc. are on hand in adequate supply. Indicated by:

- studying diagrams, reports, etc. before beginning assembly operation
- discussing work procedures with team members and lead man before beginning work



## RECOGNIZING AND SOLVING WORK PROBLEMS

Candidate is able to determine that conditions exist which may cause work problems (e.g., work stoppage due to lack of parts, quality problems (rejects), etc.). Responds to visual stimuli (e.g., information posted on boards, condition of assembly and parts, etc.) indicating work problems and requests appropriate assistance to solve the problem situation before it occurs. Indicated by:

- aware of time schedules, deadlines, etc. in performing assembly work
- repositions pieces and subassemblies to ease assembly operation
- requests clarification and/or assistance on assembly procedure from lead man

## WRITTEN COMMUNICATION SKILL

Able to read and review written assembly instructions, schematic diagrams, etc. to understand and facilitate assembly operations. Uses appropriate forms to record production information (e.g., rejects, problems, etc.). Indicated by:

- reading material which is passed "down the line"
- uses written information to answer questions, rather than relying solely on oral questions
- completes all required written forms correctly and within appropriate timeframe (e.g., quality control forms)

## PROPER ASSEMBLY AND TOOL USE

Assembling motor properly (according to schematic drawings and other instructions) to meet quality standards. Using appropriate and correct tools to facilitate the assembly operation (i.e., using the right tool for the right job). Indicated by:

- using tactile perception to fit parts together with both hands, making adjustments until proper fit achieved
- using correct tools, etc. when assembling parts and subassemblies
- positioning assembly in correct alignment to make addition of parts and/or subassemblies easier

## **WORK MOTIVATION AND INVOLVEMENT**

Maintains motivation to complete assembly task without the presence of supervisor. Displays an interest in the work being performed and takes pride in good productivity and goal attainment (team goal). Indicated by:

- keep level of work constant, even when supervisor not present
- offers to take on work assignments when needed
- makes comments indicating a desire to meet production goals

## **ORAL COMMUNICATION SKILL**

Able to speak clearly and concisely with team members and supervisors when providing information and asking questions. Understands and is able to follow oral instructions of others. Indicated by:

- phrases questions appropriately to others
- follows oral directions from supervisor and others
- provides clear and concise commentary on work process and outcomes to both team members and supervisor

## **QUALITY ORIENTATION**

Concerned with the maintenance of high level of product quality through the application of appropriate quality control procedures. Able to recognize and differentiate between good ("go") and poor ("no go") product quality. Indicated by:

- performs required quality control checks on finished subassemblies and assembly
- reviews rejects, identifying and correctly recording types of defects
- inspects work of self and others, noting quality (reject) problems

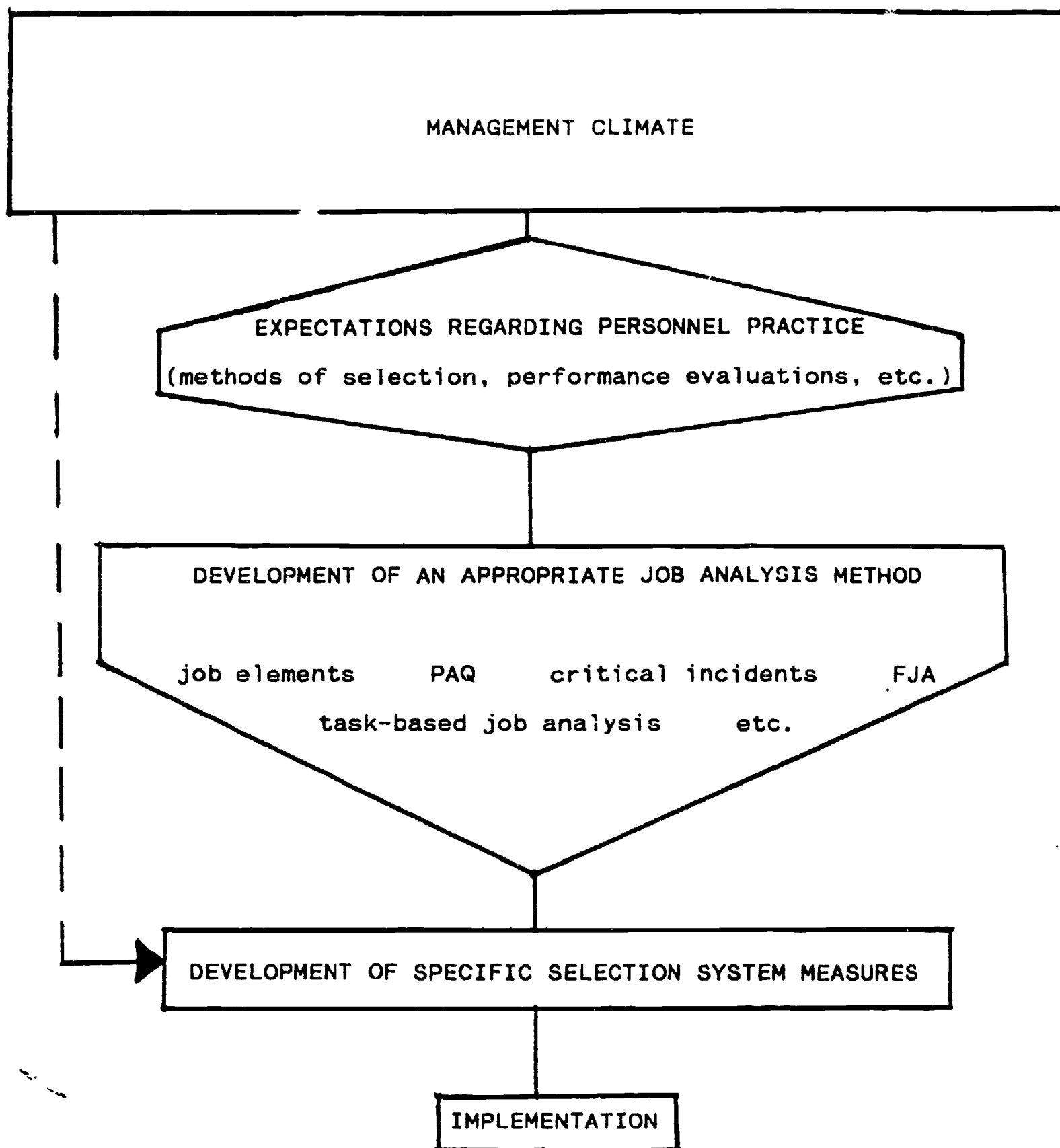


Figure 1. Decision factors in choice of job analysis method.

## SOLDERING

Able to solder correct pieces together using appropriate tools and material yielding an acceptable grade solder (based on solder appearance and strength). Indicated by:

- waiting for wire(s) to heat before soldering
- holding heat source for required three (3) count to ensure proper strength of final solder

FINDINGS AND RECOMMENDATIONS FROM A MULTI-PURPOSE  
JOB ANALYSIS OF FIRST-LEVEL SUPERVISORY CLASSES

by

Donna L. Denning

Kenneth S. Shultz

Paper presented at the International Personnel Management  
Association Assessment Council Annual Conference, San Diego,  
California, June 27, 1990.

## Part I: IMPETUS, METHODS, AND GENERAL RESULTS OF THE STUDY

Presented by Donna L. Denning

### Introduction

The City of Los Angeles has approximately 42,000 permanent, full-time employees who are classified into over 1000 different job classes. Nearly one-half of these employees are in jobs which have three or fewer incumbents; fewer than one hundred jobs have over one hundred incumbents. There are approximately thirty departments in the City, ranging in size and diversity from the four person Commission on the Status of Women to the thirty person Department of Aging to the Department of Water and Power and Police Department, each of which has over 10,000 employees.

Like most Civil Service systems, the City of Los Angeles conducts examining for entry into each job class separately. While examinations for different job classes may have some parts in common, when the jobs share some comparable requirements, the entire examination for two job classes is virtually never the same. Test content is specified within the City Charter, insofar as it stipulates that "examinations shall be practical in their character and shall relate to those matters which will fairly test the relative capacity of the persons examined to discharge the duties of the job." As such a content-oriented test construction strategy is most often seen as appropriate, although the applicability of this approach dissipates with some entry level or other jobs in which complete training will be provided. In these instances, use of aptitude testing, through the support of criterion related validation studies, is accepted as more appropriate.

In Los Angeles, examining most often consists of a multiple-choice written test and/or an interview, but these techniques may be supplemented or supplanted by an essay or technical problem-solving test, performance or physical ability test, management exercise, rating of promotability, or supplemental application blank. As a further reflection of the content-based approach to examining, credentials are emphasized; most examinations have Minimum Qualifications, in the form of educational or experience requirements, that individuals must meet to even be accepted as a candidate.

For first-level supervisory jobs, no specific educational or experience requirement must be met, but there is recognition of the need to assess what is usually termed "Supervisory Preparedness" or "Potential to Supervise." The most common means of assessing these factors has been through use of a written, multiple-choice test of supervisory job knowledge/supervisory judgment (in most instances, these two types of items are virtually indistinguishable) and in an interview. The interview



makes possible a more thorough assessment of relevant prior experience as well as, at times, providing an indication of supervisory knowledge/judgment through use of questions requiring response to hypothetical situations.

### Problem

While no particular difficulties with the present approach to the selection of first-level supervisors have been encountered, the class-by-class examination construction described, and counterpart class-by-class job analysis information on which it is based, is cumbersome and results in unexplained (and no doubt irrelevant) inconsistencies in the examining for selection into different first-level supervisory jobs. Further, on a practical note, no empirical test validation research is possible when studying one job class at a time due to the small number of incumbents in any one class. Interestingly, within the City's personnel training and performance evaluation systems, the commonality of supervisory jobs is recognized and incorporated into these systems, in that training programs provided apply to all incumbents in first-level supervisory jobs and the use of a standard performance appraisal form has long been advocated.

For these reasons, then, it was decided to conduct a comprehensive test validation study for selection into first-level supervisory jobs. Additionally, to provide enhancement for other systems and to maximize the integration of all relevant personnel systems, it was decided to include a training needs analysis and the establishment of performance appraisal criteria into the effort. Finally, due to a frequently cited need to fully communicate job information to potential candidates, and due to its Affirmative Action implications, it was decided to create a written Realistic Job Preview for prospective supervisory job candidates.

### Method

The first step in conducting this study was to identify first-level supervisory job classes. Job descriptions for all classes were reviewed, and each was coded as follows: 0=non-supervisor; 1=functional supervisor; 2=first-level supervisor; 3=second-level supervisor and above (management). Jobs selected for inclusion in this study were coded 2; there were approximately 250 such job classes with a total of approximately 4000 incumbents.

A stratified random sample, by department, of 200 incumbents in these jobs was selected for participation in the study. Participants were notified of their selection by letter. Each was scheduled into three 3-hour research test sessions. The letter also included a description of the study purpose and assured the absolute confidentiality of all research data collected.

Of central importance in this multi-purpose research effort was the job analysis. A Job Analysis Questionnaire (JAQ) was constructed from a variety of available sources: job descriptions, job analyses from previous content validation studies, and the professional literature. The final JAQ consisted of forty tasks which were rated on three scales, Frequency of performance (including a "not performed" option), Training Needed to perform, and Training Received. Finally, an estimate of the percent of the job covered by the tasks in the questionnaire was made. Other data collection instruments were a Background Information Form, to provide demographic and other descriptive information, and the Supervisory Study Questionnaire, a training needs analysis questionnaire presented in an open-ended format.

Tests included in the research were: (1) How Supervise, an older, popular short test of supervisory knowledge/judgment; (2) a specially constructed 50-item General Ability test, which used the familiar spiral omnibus format of verbal, numeric, and symbolic items; (3) a 100-item test of supervisory knowledge judgment items taken from the City item bank, grouped into six content areas; and (4) the Leatherman Leadership Questionnaire (LLQ), a recently developed 339-item test of supervisory knowledge.

Also, primarily for use in constructing the performance appraisal form, critical incidents of exceptionally good and poor supervisory job performance were collected from managers one and two levels above the first-level supervisors. Incidents were collected in four such sessions, with four to six different managers attending each.

## Results - Descriptive

### BACKGROUND INFORMATION FORM

The first "result" of this research was that too demanding a test session schedule had been established. With three consecutive week, 3-hour sessions scheduled for each participant, most participants were unable to attend all three sessions. Thus, rather than the desired 200, the sample size for each individual data collection instrument is in the 80-140 range.

In terms of demographics, the final sample attained was an average of 44 years old and had worked for the City of Los Angeles an average of 17 years. The average time in their current job was 5 years, and they had been a supervisor for an average total of 6 years. For each of these measures, there was a very wide range of values in the sample (for example, 23-67 in age and 1-28 years as a supervisor). Seventy percent of the sample was male and 30% female.

These supervisors reported directly supervising a median of 6 employees; and the sample was very near to evenly split among

those who supervise Craft/Field people (36%), Office/Clerical employees (30%), and Professional employees (34%).

#### JOB ANALYSIS QUESTIONNAIRE

All 40 tasks in the Job Analysis Questionnaire were reported performed by at least 60% of the supervisors, and a large majority of tasks were reported performed by over 70% of them. In fact, many tasks were endorsed by virtually 100% of the supervisors. Tasks in the Job Analysis Questionnaire fell into the general areas of supervisory activities in Orientation, Standards Development, Training, Plans and Assigns Work, Facilitates Subordinates Work, Reviews/Monitors Work, Performance Feedback, Performance Evaluation, Conveys Information, Performs Personnel Policy Functions, Schedules Work Hours, and Assists in Budget Preparation.

Job analysis data was also examined by type of employee supervised. This distinction was seen as particularly important, as a major concern with the design of the study was the need to combine all first-level supervisors into a single sample, regardless of the function or level of the employees they supervise. Results of this analysis indicated the extreme similarity of the supervisory job duties of those supervising Craft/Field, Office/Clerical, and Professional employees. On only eight of the forty tasks was there any notable difference in report, with most of this indicating an increased tendency for supervisors of Professional employees to do administrative/analytical work such as writing job descriptions and performance standards and budgeting. Finally, the average portion of their job covered by the tasks was reported to be approximately 70%, and this did not differ among the three groups of supervisors.

On the training rating scales, the average training needed on most tasks corresponded to the scale value "a lot," while the average training received was typically reported to be "a little." Nevertheless, the correlation between these two scales was very high ( $r=.90$ ;  $p<.0001$ ), indicating that, while incumbents report needing more training than they are receiving, they also perceive that they receive relatively more training in those areas in which they need it most.

#### SUPERVISORY STUDY QUESTIONNAIRE

When asked, in an open-ended format, the main weakness of City supervisors in general, responses tended to be quite broad and vague: "general supervision," "attitude," and the like. When asked about personal weakness, however, both in terms of current training needed and difficulties when first promoted, responses were more specific. Supervisors reported that their greatest need is in the area of evaluating and disciplining employees.

## HOW SUPERVISE

The average score on this 70-item test was 47%, and the three groups of supervisors did not differ significantly in their average scores.

## GENERAL ABILITY

The average scores on this test were 50%, 40%, and 36% for the incumbents who supervise Professional, Office/Clerical, and Craft/Field employees, respectively. The highest scoring group scored significantly higher than the two lower scoring groups, which did not score significantly different from each other.

## ITEM BANK

The average scores were 79% for the supervisors of Professional employees, 68% for supervisors of Office/Clerical employees, and 66% for supervisors of Craft/Field employees. Again, the supervisors of Professional employees scored significantly higher than the other two groups, which did not differ from each other.

## LEATHERMAN LEADERSHIP QUESTIONNAIRE

Supervisors of Professional employees scored an average of 65%, which was significantly higher than the Office/Clerical supervisors, who scored an average of 53%, and the supervisors of Craft/Field employees, whose average score was 52%.

## PERFORMANCE APPRAISALS

Two levels of supervision above the first-level supervisors in this study were requested to independently complete a performance appraisal form for each study participant. While it was not possible to get both appraisals in all cases, at least one appraisal was attained in almost every case.

On average, the supervisors of Office/Clerical employees were rated highest by their superiors (6.57 on the 9 point scale), supervisors of Professional employees were rated an average of 6.10, and supervisors of Craft/Field employees were rated an average of 5.67. In this case, the highest rated group's rating is significantly higher than the lowest rated group's rating, but the middle group is not significantly different from the other two.



## Results - Substantive

### REALISTIC JOB PREVIEW

From all sources of job information, the Job Analysis Questionnaire, the Supervisory Study Questionnaire, and the critical incidents, a Realistic Job Preview was written. It emphasizes the large amount of paperwork required of supervisors, the interpersonal demands of the job, and the need to get work done through others without doing it personally. As is typical of Realistic Job Previews, there is no attempt to provide a comprehensive description of the job; rather, it highlights the often less obvious aspects of the job of a first-level supervisor. A copy of the Realistic Job preview is included as Appendix A.

### TEST VALIDATION

Data analyzed for the test validation portion of this study were: the How Supervisor and General Ability test scores, the six Item Bank scores and an overall IB score, the twenty-seven LLQ scores and an overall LLQ score, and job performance ratings. For job performance, ratings provided by the first- and second-level managers were evaluated separately. While they demonstrated adequate reliability (intercorrelations of separate factors were mostly in the .40 - .60 range), there were far fewer second-level ratings obtained, so to rely on the combined ratings would have drastically reduced the sample size. A composite rating was created for each rater by averaging across all factors.

Both the How Supervise and General Ability test failed to yield more than a chance number of statistically significant correlations with the 19 performance ratings. While all correlations were positive, a large proportion of them were very near zero. This is surprising, particularly in light of the recent evidence of generalized validity of cognitive ability tests, especially in jobs with significant information processing demands. Beyond citing sampling error, a possible reason for this result in the current research is that the General Ability Test demonstrated the largest differential in scores among the three supervisory groups. No such differential was demonstrated for the job performance ratings.

On the Item Bank test, a consistent pattern of positive correlation with first-level manager ratings was attained, with five of the 19 correlations achieving statistical significance (at least  $p < .05$ ) and an additional eight demonstrating marginal ( $p < .10$ ) significance. The correlation with the average job performance rating was .20 ( $p = .05$ ).

Looking at the correlations of the six separate scores on the Item Bank test with average job performance, two scales yielded statistically significant correlations (Evaluating Employees,  $p < .01$ ; and Motivation and Development,  $p < .05$ ), two were marginally significant (Discipline and Training), and two were correlated near zero (Planning and Assigning Work and EEO/AA Responsibility).

Correlations with each first-level supervisory performance rating factor were also calculated for the total Leatherman Leadership Questionnaire score. These correlations ranged from .13 to .35, with fourteen of the nineteen correlations achieving statistical significance. As with the Item Bank test, the twenty-seven subscales yielded considerably different results when individually correlated with job performance. They ranged from .13 to .43, with the single exception of one value of  $-.06$ . In all, fifteen of the twenty-seven separate correlations were statistically significant.

The same series of correlations was computed with second-level manager ratings, which yielded much the same pattern of results, although far fewer statistically significant correlations were attained. Again, however, it was the Item Bank and LLQ tests for which statistical validity was indicated.

#### TRAINING NEEDS ANALYSIS

In combination, the three training needs analysis data collection sources, the Job Analysis Questionnaire, the Supervisory Study Questionnaire, and the Leatherman Leadership Questionnaire consistently point to training need in the areas of handling employee grievances and disciplinary problems, evaluating employees, and conducting selection interviews. Beyond these areas, there was much diversity in response to the question of training needs, ranging from numerous indications of need in communication skills to technical job skills to planning and scheduling work and conducting on-the-job training for employees.

#### PERFORMANCE APPRAISAL

As mentioned previously, a performance appraisal form for research purposes was constructed for use in the current investigation. This form was devised primarily from information attained in critical incident sessions with managers. Data from the Job Analysis Questionnaire was also used as a supplement and cross-check to the performance factors identified. See Appendix B.

A form for operational use was adapted from the research form. The only alterations necessary proved to be in format, as the research form was accepted as reflecting the important aspects of supervision on which evaluation and feedback should take place for developmental and administrative purposes.



## Part II: RECOMMENDATIONS AND IMPLEMENTATION OF THE STUDY RESULTS

Presented by Kenneth S. Shultz

The first recommendation from this study is that results be applied to all first-level supervisory classes. Recent research on the topic of validity generalization/transportability has demonstrated that when jobs are "substantially similar" to results from a study on one job may be applied to other substantially similar jobs. To determine the appropriateness of applying the results from this study to any individual job class, the Job Analysis Questionnaire should be completed by a sample of job experts (e.g., supervisors and/or incumbents) for that class, and results of this analysis should indicate the substantial similarity of the job under consideration with the jobs in the original study to indicate the "transportability" of results.

For example, the need for an examination for a given first-level supervisory job class (e.g., Senior Plumber) would be established by the head of the appropriate examining section. An analyst in the Test Research Section would then administer the Job Analysis Questionnaire (containing only the "essential" tasks -- see previous discussion) from this study to a sample of incumbents in the class of Senior Plumber. If results from the Senior Plumber analysis showed 70% of the tasks to be essential, then the results of this study would be transportable to the Senior Plumber class.

Once job analysis has verified the essential similarity of job activities, all study results should be applied to the job class in question. The Realistic Job Preview (RJP) is presently being reproduced by our personnel training section using a desk top publishing program. We will attempt to distribute the RJP to all potential candidates for entry into first-level supervisory classes. One avenue to accomplish this will be to place a copy of the RJP in the "Training Bulletin" published periodically by the Training Section of the Personnel Department. In addition, the RJP will be disseminated in appropriate training classes such as "Introduction to Supervision" and "Career Development". The RJP will also be available to the department personnel offices as well as to the Recruitment Division for individuals who pick up an application for a first-level supervisory class.

Next, examining for first-level supervisory jobs should proceed along the lines indicated in this study, namely use of a 50% weighted interview and a 50% weighted multiple choice test is typically warranted, although an alternative to this weighting scheme may sometimes be appropriate (e.g., 100% interview). Continued use of content-based supervisory knowledge/judgment multiple choice items in the written test is appropriate based on the results reported earlier. While the failure of the General Ability test to demonstrate statistical validity in this study

should certainly not be overinterpreted, results clearly indicate the acceptability of continued use of content-based supervisory items. For purposes of test security, many such items will need to be available, as the exact test must change for different examination administrations. To provide the best possible tests within this constraint, approximately 100 new items are currently being written and will be evaluated (e.g., subjected to item analysis) as they are used in different examinations.

In addition, the 1,000 plus items on supervision in the Personnel Department's item bank are being reviewed to ensure that the items are not contradictory to the findings of the present research. For example, some items may reflect earlier views on supervision that are no longer accepted as desirable (e.g., the best way to motivate all employees is with money). The categorization of items is also being reviewed, to the extent possible, to make sure that individual items are properly placed as well as to make sure the overall categorization scheme is in line with the results of the present study. The categories that are most in need of new items will be the focus of the new item writing mentioned above.

The need to continue use of an interview as one component of supervisory selection is clearly indicated from the job analysis given the extreme oral communication and interpersonal demands of these jobs. A "Resource Guide for Constructing First-Level Supervisory Interviews" has been drafted and when finalized will serve as a reference to examination analysts developing interviews for first-level supervisory job classes. Interview factors in the guide are identified and defined through use of a behavioral consistency model in which interview factors are based on the job performance dimensions determined from job analysis information. Relevant behavior-based questions accompany each of the interview factors. It is important to note that behavioral consistency refers to consistency in relevant and related behaviors and is not simply limited to actual samplings of job behaviors. This is an important distinction, given that many applicants for first-level supervisory jobs may have no actual experience in doing supervisory tasks, therefore it is their preparedness or potential to supervise that should be evaluated. In other words, absent any direct relevant experience, we are looking for "signs" of who would make an effective first-level supervisor, rather than actual "samples" of previous first-level supervisory work.

This guide is being integrated with the new edition of the "Examining Division Procedural Manual" which details all aspects of the civil service selection process including discussions of topics such as types of interviews, how to set up interview boards, and developing interview scoresheets. Appropriate interview content for a particular job class will be determined from the information in the comprehensive job analysis for that particular class and in conjunction with the transportability information.

Implementation of the Training Needs Analysis results is largely an exercise in integrating these findings into existing supervisory training programs. There are already several such training programs available on a City-wide basis (e.g., Introduction to Supervision, Discipline and the Supervisor, Effective Leadership) so enhancements and areas in need of increased emphasis within these existing training programs are basically what this study highlights. Additionally, it is recognized that these areas are identified within the study only in very general terms. To enrich these descriptions, and provide further direction for training program modifications, more detailed information will be obtained in a survey distributed to all new (i.e., appointed since July 1, 1989) incumbent first-level supervisors. The survey will ask these new supervisors about their training needs in a very systematic way (e.g., using multiple choice questions as opposed to the open ended questions used in the present study).

Presently, the supervisory development courses have many more applicants than openings available. For example, three courses offered -- "Certification Interview Training", "Discipline and the Supervisor", and "Effective Leadership" -- had over 1,600 people apply for the just over 400 slots available in the January-June, 1990 training classes. Because of the large discrepancy between the number of new supervisors desiring training and the number of slots available in training, a significant time lag can occur between when a new supervisor starts and the time when they received the needed training. In order to fill this void a "Supervisors Survival Kit" has been proposed. A sample table of contents describes topics such as managing skills (e.g., delegating, discipline, time management), administrative issues (e.g., absenteeism, insubordination, staffing, training), and carrying out City policy (e.g., documentation, EEO/AA, sexual harassment). The kit will serve as a guide to new supervisors by supplying them with pragmatic general information on supervision and leadership as well as practical information on specific city procedures such as what to do in a given situation (e.g., what form to fill out when an employee is absent) or where to find needed information.

It is important to note that the kit is being developed to help the new supervisor in those confusing and anxious first few weeks or months before they can obtain the training mentioned above. The kit is in no way intended to supplant training, rather it is simply filling in during the interval between when the individual becomes a new supervisor and when they can be scheduled for training. The information in the kit is fully in line with the information provided in training and will be expanded and explained more fully in the actual training sessions.

Traditionally, the use of a formal performance appraisal system in many City departments is not consistently present. It

appears that, in many instance, employees receive only sporadic feedback regarding their performance, especially after passing probation. This lack of formal appraisal lead the City Council to order the Personnel Department to develop a new performance appraisal process. The result was the "Supervisor's Guide to Performance Appraisal" -- a handbook which discusses topics such as how to set standards and hold the appraisal interview, and which also contains sample performance appraisal forms. An operational version of the Supervisory Performance Appraisal form will also be included in future additions of the guide. The guide will help supervisors better understand the performance appraisal process and will hopefully allow them to be more willing to use such a process.

In conclusion, the efficacy of a multi-purpose job analysis approach has been demonstrated. In this case, the "multi-purpose" was designed into the study in two ways. First, Each data collection instrument did (at least) double duty in serving more than one purpose. Examples of the multiple uses of data collection instruments include the use of the Leatherman Leadership Questionnaire for purposes of test validation and training needs analysis and the use of critical incident information for performance appraisal form construction, Realistic Job Preview creation, and as a cross-check on the job analysis data. The second "multi-purpose" was the several outcomes of the study, which included test validation, training needs assessment, development of a realistic job preview, and development of a performance appraisal form. By researching these issues in a single study, the intrinsic interrelationships between these seemingly different personnel functions are accentuated and the ability to use the same general information for several purposes is highlighted.

# **A CAREER DEVELOPMENT ASSESSMENT CENTER FOR COURT MANAGER**

**Issues and Opportunities**

**Patrick T. Maher  
Principal Associate**

**A presentation to the 1990 International Personnel Management Association Assessment Council  
Conference on Personnel Assessment**



**Personnel and Organization Development Consultants, Inc.**

5842 Crocus Circle La Palma, California 90623 Phone (714) 827-1780



The Municipal Court, Los Angeles Judicial District (LAMC) have had a long and distinguished tradition of providing extensive in-service training for all of its personnel. Among the many training courses provided are courses designed to teach necessary skills and abilities for supervisors and managers.

One of the problems inherent in developing any training program, however, is determining what training is necessary to overcome deficiencies in current supervisors and managers. While general training programs may assist in developing and providing overall skills, they often do not target the specific needs of individuals. Thus, a generally-applied training program may provide training to an individual who already has proficiency in that area, while, at the same time, not provide training in an area in which the individual needs to improve performance.

To overcome this training problem and identify and satisfy individuals' specific needs, LAMC elected to institute a career-development program. Not only does a career-development program identify and correct individual deficiencies for those in their current assignments, it also helps develop personnel and prepare them for advancement to higher levels. Such programs provide a return to the individual who participates in them, by preparing him/her for advancement, while at the same time assisting the organization by ensuring that candidates for promotions are highly qualified, experienced and ready for advancement.

To accomplish these goals, LAMC decided to initially institute a career development process for its court manager positions. In LAMC, the court manager operates as a second-line supervisor/middle manager. Because of the number of court managers, it was deemed appropriate to start with this level as a means of preparing individuals for advancement into management, as it would provide a large-enough pool of individuals to have a definite impact on the overall organization.

Executive staff eventually elected to use an assessment center process, after coming to the conclusion that the assessment center process provided the best opportunity to assess performance within the context of actual job duties as well as obtain meaningful information that would permit the identification of individuals' developmental needs. The responsibility for implementing the career-development process and the assessment center as a part of that career-development process fell to the training office. The training office was selected for this responsibility because, ultimately, the training staff would be responsible for developing both court and individual training programs. Thus, it seemed most logical to place the responsibility for the entire process with that unit.

Initially, a job analysis was conducted for the position, which was newly created by combining duties of two other positions that would be eliminated through attrition. The position of court manager was, for all practical purposes, a new one.

This created some problems, because new job specifications were not yet complete, and there was not complete agreement among incumbents and supervisors as to what specific duties were to be allocated to the classification. Thus, the job analysis served, to a great extent, to identify such duties.

However, some problems were encountered. There was agreement that the incumbents should perform certain duties, but the processes for performing such duties had not been implemented. The most notable example of this was the responsibility to monitor budgetary expenditures within the unit of assignment. In this case, both incumbents and their supervisors felt it was a critical task, but the means to obtain timely expenditure reports were not operational.

In a few other cases, there was not sufficient consensus among incumbents and their supervisors as to what extent some duties should be performed. Thus, it was difficult to construct an assessment procedure to assess all skills related to all critical tasks. However, there was sufficient agreement on which tasks and the skills to perform them were to allow a sufficient career-development assessment, if not a selection assessment.

Since the purpose of the assessment procedure was to evaluate performance strengths and limitations, there was no need to develop a rank order list. Thus, there was no time spent on determining whether or not critical skills differentiated in performance, such as would be necessary for a promotional or hiring process requiring a rank-ordered list.

Generally, however, the content validation process followed the one necessary for validation of a selection process.

One of the early realizations was that the assessment center process is labor-intensive, in that a minimum staff of four individuals is required to assess six participants in the assessment center.



Furthermore, a minimum of two days is required for the assessment center staff and one day required for the participants. Thus, it became readily apparent that not all court managers could be initially placed through a full assessment center process, due to both staff and budgetary limitations.

However, it was also recognized that, unless some type of career-development activity were to take place for the majority, if not all, of the court managers, a career-development process limited to only a very small number of individuals would have little organizational value. Fortunately, LAMC was able to institute procedures that met the need of providing career-development analyses and feedback to a large number of individuals for a minimal cost.

The General Management In-Basket (GMIB) was used as a screening tool to select the most qualified group of court managers. The GMIB is a relatively inexpensive, commercially available in-basket test that measures generic supervisory/management skills. It has been empirically validated and has a national database with norms for performance on four assessment dimensions: (1) leadership, (2) handling priorities, (3) managing conflict, and (4) management control. Importantly, GMIB reports are useful not only for selection but for developmental purposes.

Because of the ease of administration and the cost effectiveness of the GMIB, senior court managers were also given the opportunity to take the GMIB. This broadened the number of personnel who were able to participate in career-development activities.

All participants who took the GMIB received feedback on their performance. This fulfilled the need to provide some developmental assistance to the entire candidate group. The feedback consisted of: (1) a bar chart profile depicting how the candidate scored relative to the national database on each of four assessment dimensions measured by the GMIB; and (2) a GMIB Candidate Feedback Report describing candidate performance in each of the four assessment dimensions. Thus, candidates were given feedback on their current skill level in four important assessment dimensions compared to other supervisors and managers, as well as written reports highlighting their strengths and weaknesses.

Once the results of the GMIB were obtained, LAMC selected six persons from among the top ten court managers based on the GMIB results. These six court managers then participated in the rest of the assessment center process.

It should be noted that this was simply a means of selecting the first participants in the process. It is anticipated that other court managers will be participating in the assessment center since it will be an ongoing process.

Critical to any assessment center process is the need for **trained assessors**. This need was met by using assessors who had trained for three days in the assessment center process and then gone through a one day certification process in which they demonstrated their assessor skills (an assessment center for assessors). This training meets or exceeds that required in the **1989 Guidelines and Ethical Considerations for Assessment Center Operations (Guidelines)**.

Three members of the staff of the municipal court went through the training and served as assessors. However, to ensure greater acceptance of the process's objectivity, increase ethnic diversity, and assist the newly-trained assessors in their first assessment center, three other trained and experienced assessors were obtained on a reciprocal basis from other agencies. Thus, a team of six assessors were used the first time, even though only three assessors were required.

The assessment center process used consisted of four common assessment center exercises. First, there was an assigned-role leaderless group discussion. In this exercise, six participants are assigned a task and given a period of time (in this case, one hour) to come to an agreement within the group of how the situation should be solved.

This exercise was combined with two others, an oral presentation, and a written problem exercise. As part of preparation for the group discussion, the candidates were also given time to prepare a formal presentation to other group members on their respective assigned position. They were provided an overhead projector with transparencies and a flip chart, and encouraged to use these visual aids as a part of their oral presentation. This allowed assessors to determine their formal presentation skills. Subsequent to completion of the group discussion and the group arriving at a final decision, the candidates were told to prepare a memorandum explaining the basis for the group's decision. Collectively, this combination of exercises, based on the same scenario, permitted the analysis or evaluation of a number of assessment dimensions.

The final exercise involved an interview simulation. The candidates were provided with some background information on a specific employee and given time to become familiar with this information.

They were then instructed to interview the employee and make a final determination as to what the problems were and how they were going to deal with them in the future.

Once the participants had completed all of the exercises, the assessors met for purposes of determining performance levels, within dimensions, across all the exercises. They transferred documented behavior from the notes taken during their observation of the participants in the exercise to the evaluation forms and shared this information among each other (an assessment center process commonly referred to as consensus or assessor discussion).

Although assessors provided scores for every candidate within each exercise and each dimension, overall scores were not calculated. Furthermore, time was not spent attempting to obtain consensus among assessors where scores were different. Instead, scores were used to focus on significant behaviors, for feedback to the candidates. Thus, scores were not reliable indicators of relative performance for selection purposes.

Once all candidates had been evaluated in this process, the assessment center administrator prepared summary feedback reports identifying overall performance within exercises. The administrator also identified primary developmental needs, based upon input obtained from the assessors during assessor discussion. Certain developmental recommendations were also made for the individuals.

In some instances, all participants exhibited the same developmental needs, and, in these cases, all participants received the same basic developmental recommendations. This information can also be utilized by the training staff to develop general training programs for all court managers.

The training staff also retained a copy of the developmental feedback reports, in order to assist individual court managers in identifying training programs or experiences that would help meet the developmental needs identified through the assessment center process.

Once the feedback reports were completed, the assessment center administrator met with each participant and personally reviewed the feedback reports and assessment center results. This personal interaction allowed the participants to both obtain clarification on their individual performance and explore in more detail developmental needs and recommendations.

Overall, the candidates who went through the entire process seemed to think that it was beneficial and job related. Although not evaluated through any criterion process, candidates generally agreed that the assessment center process accurately pinpointed their strengths and weaknesses.

The lesson to be learned here, however, is that, whether or not an assessment center is to be used for career-development or selection purposes, it is imperative that it be done correctly. Too many shortcuts or too little commitment to the process, especially budgetary commitment, will likely doom the entire project. Strict adherence to the Guidelines, not only in training, but in all aspects of the assessment center process is essential for any assessment process to be successful.



# County of San Diego

ETHEL M. CHASTAIN  
DIRECTOR

## DEPARTMENT OF HUMAN RESOURCES

1600 PACIFIC HIGHWAY ROOM 207 SAN DIEGO CALIFORNIA 92101 2483

### SAN DIEGO COUNTY MANAGEMENT ACADEMY

#### A Program that Succeeded

April 9, 1990

Two years ago I came to the IMPAAC Conference in Las Vegas with a description of the San Diego County Management Academy which had been conceived and implemented using assessment center methodology for selection. At that time we had no data to demonstrate that the process which we were using was either effective in achieving our goals or acceptable to the participants or executives within the County. Today, I can report to you that we're doing what we said we were going to do, and, as a result, we're changing the ways in which County executives are making appointments to positions of responsibility.

For those who did not hear the presentation two years ago, as well as others who are unaware of the San Diego Management Academy, I'll briefly summarize the program for you.

Conceived in 1985, the Management Academy is both a productivity and an affirmative action program. A small technical team was assembled early in 1986 to flesh out the program concept and model it for County executives and interest groups. The first year was devoted to research into management development programs and selection instruments, the modeling of alternatives, and the final selection of a program design.

Although competitive in nature, the program is developmental for permanent County employees. There are no guarantees of promotion, but as you'll see later, promotion has come to many of the participants. Recruitment is aggressive, intended to reach all County employees, and interest is keen. Essentially, there are only two application criteria: first, the applicant must be a permanent County employee who has passed probation in any class, and second, the employee must be rated standard or higher in all rating factors on the latest performance evaluation.

Selection is by competition utilizing work simulations. A consultant was used to design our assessment center to ensure that it met the Guidelines and Ethical Considerations as established by International Congress on Assessment Center Operations. The preliminary screening instrument is an in-basket exercise and we've used this instrument on groups ranging in size from 156 to 403 participants. Applicant groups of reduced size then participate in unassigned role leaderless group discussions, written and oral reports, and subordinate counseling exercises. Raters are generally County executives who have all been trained by the design consultant with three or more days of formal training. Since its inception in 1987, 1801 have employees have applied for the program, 1271 have participated in the in-basket exercises, 217 in the assessment centers, and 123 have been selected as management candidates.

The final selection process consists of two phases; the setting of a minimum pass point on the assessment center simulations, then selection by ethnic group with the condition that the total of minority selections should be at least equal to the total of the caucasian selections. This selection criteria has served us well.

Once selected, management candidates receive extensive feedback on their assessment center performance, negotiate an individual development plan (usually with a senior executive within their own department), attend familiarization classes in those areas where they have little or no experience, participate in formal training, and undertake special projects. These plans are extensive and ordinarily take from 18 to 30 months to complete.

To measure effectiveness of the program we established control groups for each Academy class and we compare the promotion rate of the control group to that of the Academy class. The members of each control group are matched to individual management candidates in classification, length of service with the County, length of service in class, age, and ethnic group insofar as possible. The most recent comparison of promotion rate is as follows:

<u>Group</u>	Combined Group	Minorities	Caucasians
<u>Promotion Rate for</u>			
Management Candidates	52.3%	59.4%	45.5%
Control Group	23.1%	17.6%	26.6%

As you can see, management candidates are considerably more upwardly mobile than their control group counterparts. Not counted in the statistics are the multiple promotions. Several candidates have been promoted three times since notified of their selection for the Academy. Others have been promoted twice. Department



heads are more than eager to take advantage of those individuals who demonstrate effective management and supervision skills, skills which we evaluate in the selection process for the Academy.

The Academy has gained an excellent reputation in its short existence. Most of those executives who were doubters when we began in 1987 are now strong supporters. A very few still remain to be won over. Nonetheless, employees from all departments are convinced that the Academy is a stepping stone to a more mobile and satisfying career in County government and are applying and gaining admittance. I believe that the Academy has gained a permanent place in the County scheme -- a credit to current candidates and graduates alike.

Of great interest is the side effect which the Academy selection process has had on County government. As noted earlier, a large percentage of our assessors are at the executive and senior management level; department directors, assistant and deputy directors, and division managers. They've seen the value of management simulations and are asking for and using them in the departmental selection process. In the recent past we used in-baskets for Supervising Probation Officer and Sheriff's Captain, and more extensive work simulations for Deputy Chief Probation Officer, two Assistant Deputy Directors in General Services, Risk Manager, Employee Benefits Manager, Social Services Administrators I and II, Supervising Eligibility Technician, and Court Services supervisors and administrators.

In the County of San Diego we've found a way to identify the up and coming employees, to provide them with the visibility and networking to become effective, and promoting them to more responsible positions. Today's management candidates are tomorrow's executives and we're getting them ready for future responsibilities.

Prepared by:

Del Boerner  
Personnel Services Manager  
(619) 531-5140

A Follow-up to the 1989 Criterion Related Validity Study  
Conducted to Select Lieutenants in the  
Palm Beach County Fire Rescue Department

Linsey Craig  
Manager, Recruitment & Selection  
Board of County Commissioners  
Department of Employee Relations & Personnel  
West Palm Beach, Florida  
Paper presented at the 1990 International Personnel Management  
Assessment Council Annual Conference  
June 24-28, San Diego, CA.



### Background

The use of assessment centers in the public sector has been prevalent since the 1970's (Fitzgerald, 1980; Fitzgerald & Quaintance, 1982; Mendoza & Craig, 1983; Joiner, 1984; and Yeager, 1986). The fire service is one profession that has used the process widely. In fact, several landmark U.S. courts cases pertinent to personnel selection including assessment centers (Firefighters Institute for Racial Equality v. City of St. Louis) involved Fire service promotions (see Byham, Review of Legal Cases and Opinions dealing with Assessment Centers and Content Validity, 1983). Nonetheless, at the time of this study it was not known whether any of the Fire Departments had conducted criterion related validity studies, and if they had, what their results were.

In 1986 the Palm Beach County Fire-Rescue and Employee Relations & Personnel Departments worked cooperatively to design, develop and implement three content valid selection processes which included the assessment center for the ranks of District Chief, Captain and Lieutenant. Implementation of the assessment center process in Palm Beach County is noteworthy for several reasons. First, until 1984, Palm Beach County used traditional selection techniques to fill vacant positions. Second, 10 separate fire taxing districts representing 10 separate departments were consolidated, so many decisions including those pertinent to personnel selection had to be made in a relatively short period of time. Third, most of the districts had utilized a written test and then an appointment process that gave the Chiefs great latitude to select whom they wanted and were fearful of changes to their systems. Nonetheless, in early 1985 after extensive contract negotiations, the assessment center process became a component of Article 18 of the IAFF contract less than 2 years after the department was formed. Union representatives and management agreed that the assessment center as presented to them by Personnel, was a fair process that had a proven track record for criterion validity.

A year after the promotions were made, the department was quite satisfied with the job performance of the various personnel who had been promoted. However, statistical analyses to establish criterion validity had not been conducted. The 1986 Fire Lieutenant assessment center process is the subject of this follow-up criterion related validity study.

### Assessment Center Design and Development

The first step in the process was completion of a thorough job analysis. A modified version of the Retrospective Critical Incident Method was utilized. This method was recommended by Dr. Larry O'Leary who served as the principal consultant. After review of pertinent documents which included existing job descriptions and organization charts, a series of on-site job visits to a number of fire stations was made and standard 1/2 hour interviews were conducted with 10 Fire Lieutenants. After analysis of this information, task and competency lists were developed (see Appendices 1 & 2). The ten Lieutenants rated and ranked the competencies and provided a minimum of two critical incidents for

each task. The job analysis information was summarized and then verified by three Chief Fire Officers and the Fire-Rescue Administrator. A written exam was also developed and was the first hurdle in the selection process. Three exercises were developed; a Leaderless Group Discussion (L.G.D.), a Coaching/Counseling and a Citizen Complaint Encounter. A competency exercise grid indicated the equal weight each assessment center skill dimension was given in each exercise based upon the job analysis.

#### Method

Thirty-eight candidates participated in the 1986 assessment center. Sixteen Fire-Rescue personnel holding the rank of Lieutenant or Captain from other fire jurisdictions within the state of Florida were trained as assessors. Assessor training lasted three 8 hour days. The next day, 38 candidates participated in the assessment center. Team consensus meetings occurred the following day. A one year eligibility list comprised of 32 candidates was established. From this list 18 personnel were promoted to the rank of Fire Lieutenant and are the subjects of this study.

Thirty-four variables were analyzed and are listed in Table 1. The criteria were the standard Firefighter Performance Appraisal ratings for 4 rating periods after the promotion date; 4 months, 1 year, 2 years, and 3 years later. The Palm Beach County performance appraisal form evaluates an employee based on five categories from a high of "exceeds job requirements" to a low of "unsatisfactory" on nine performance factors (i.e., (abbreviated: coordinates and directs at all fire scenes/emergencies; evaluates and reviews subordinate performance; promotes employee training and development; resolves disciplinary problems and makes recommendations; resolves employee complaints/grievances/problems; exchanges information with superiors/subordinates; prepares reports/paperwork; accepts responsibility for decisions; encourages teamwork). So, each performance rating verbal descriptor was converted to a rating scale of 1-5 for statistical analysis purposes.

Additional criteria (i.e., the fifth set of performance ratings) included the performance ratings for each of the eight assessment center dimensions and an overall performance rating for each Fire Lieutenant. These data were obtained during interviews with the Fire District Chiefs or Captains who supervised the Lieutenants as of November, 1989. Most had previously rated the Lieutenants. The data were collected using a comprehensive form that included the assessment center skill dimensions, performance factors, two rating scales and interviewee background information. The Chiefs and Captains were told that the ratings would have no effect whatsoever on the individual but were being collected for research purposes only. Overall, 5 different performance ratings were obtained for each Lieutenant since September, 1986. During the interviews, ratings were also obtained on the 9 performance appraisal factors previously referred to.

#### Analyses and Results

The means and standard deviations for all variables analyzed in

this section are displayed in Table 1. The maximum number of applicants analyzed is 37 not 38. Data on one applicant was not complete enough to use for these analyses. The statistical package STATPAC GOLD, was used for the analyses to be described below.

#### Relationship Among Exercises and Between Exercises and the Assessment Center Total Score

Ratings were summed over dimensions for each exercise each applicant. The resulting sums were inter-correlated among exercises. In addition, the sums for each exercise were correlated with the assessment center total score. The results are shown in Table 2. As displayed, the correlations among the 3 exercises are all substantial and significant. As would be expected, the relationship between each exercise and the assessment center total score is substantial and significant.

#### Relationship Among Dimensions and Between Dimensions and Relationship Between Dimensions and Assessment Center Total Score

Ratings on each dimension were averaged over assessors for each applicant to generate a mean score for each individual for each of the eight dimensions. The 8 dimensions were then correlated to produce the results shown in Table 3. As would be expected, the intercorrelations are all positive. In addition, the more complex dimensions, such as Judgement and Decisiveness have significant correlations with most of the other dimensions. Table 4 shows the relationship between each dimension and the total score. All the dimensions have significant and substantial correlations with the total score.

#### Relationship Among 1986 Assessment Center Ratings and 1989 Supervisory Ratings on Assessment Center Dimensions

Ratings on each assessment center dimension were intercorrelated with the supervisory ratings on the 8 assessment center dimensions to produce the results shown in Table 5. The intercorrelations between the ratings for the same dimension appear in the diagonal of the table. It might be expected that these intercorrelations would be the highest for the row and column with the same dimension label. Except for the diagonal, this table only shows those intercorrelations which were .20 or higher. The expected pattern did not occur. Most of the intercorrelations were relatively small and nonsignificant. A key difference between the ratings was the very different circumstances under which the two sets of ratings occurred.

#### Relationship Between Supervisory Ratings on Assessment Center Dimensions and Supervisory Ratings on Performance Appraisal Factors

Supervisory ratings on each assessment center dimension for each applicant were intercorrelated with the supervisory ratings on the 9 performance appraisal factors. The intercorrelations were all positive and the majority were significantly different from zero. Nonetheless, certain factors were expected to have higher correlations with some skill dimensions as compared to others. For example, it was not expected that preparing reports/paperwork would be as highly correlated with dealing with people, and development



of subordinates as it was with planning and organization and such was the case as shown in Table 6.

#### Performance Measures

#### Relationship Between Assessment Center Total Score and 1, 2, and 3 Year Performance Evaluations

Of the 37 applicants, 18 were hired based on their assessment center total score. Assessment center total scores for individuals were compared to several performance measures for those individuals. First of all, average ratings on the exercises for individuals were compared to the operational performance appraisal ratings conducted at 4 months, 1 year, 2 years, and 3 years after hiring. As stated earlier, a fifth set of performance ratings were also obtained from interviews conducted with each supervisor.

One exercise, Coaching and Counseling, did obtain substantial correlations with the 1, 2, and 3 year performance evaluation (i.e., .27, .26 and .38). Only one dimension, Dealing with People obtained a significant correlation with the 1 year performance evaluation. The assessment center total score did not significantly correlate with the 4 month performance rating ( $r = .11$ ). However, as shown in Table 7, the correlations between the total assessment center score and the 1, 2, and 3 year performance evaluations were positive and substantial (i.e., .35, .36, and .40) although not significant. The probabilities are shown in the parentheses. Given the small N, the lack of significance is not surprising. These coefficients were corrected for restriction of range on the predictor since the assessment center total score was used to select the hirees. The corrected correlation coefficients are also shown in Table 7.

It is important to note that a substantially lower correlation was obtained for the 4 month performance ratings. The standard deviation for these ratings is almost double the standard deviation for the 1, 2 and 3 year performance ratings. This might mean that 4 months is too short of a time for an accurate appraisal or that the raters had different approaches among themselves during this rating period.

#### Relationship Between Assessment Center Total Score and Supervisory Ratings on Skill Dimension and Overall Performance (Fifth Set of Performance Data)

The fifth set of performance measures consisting of supervisory ratings on the assessment center dimensions and the overall performance rating for each incumbent and the assessment total score were also collected and analyzed. These ratings were collected over three years after promotion during interviews with the supervisors. The supervisors were told that the ratings were for research purposes only and would be kept strictly confidential. The correlations between these measures and the assessment center total score are shown in Table 8. As can be seen, all the correlations are positive and substantial. These coefficients were also corrected for restriction of range as displayed in this table.

#### Relationship Between 1986 Assessment Center Total Score and 1989 Supervisory Ratings on Performance Appraisal Factors

The 1986 assessment center total scores for each successful

applicant were correlated with the 1989 supervisory ratings on the 9 performance appraisal factors to produce the results shown in Table 9. As can be seen 4 of 9 are significant. Interestingly, these four performance factors all concern Dealing with People: the same dimension that exists in the Coaching and Counseling exercise and the Dealing with People dimension both which had relatively strong relationships to the performance evaluations. Although the correlation coefficients are higher when the supervisors rated the employees on these factors than on the 8 assessment center dimensions (see Table 8), this is not surprising. They had previously rated these employees using the same factors but had never rated them using the 8 assessment center dimensions. Four of these coefficients obtain significance while the coefficients based on the 1, 2, and 3 year evaluations only approach significance. This may be partly due to the fact that the 1989 ratings were collected with the knowledge that they were for research purposes only.

### Conclusions

The results of this study should be viewed within the context of prior research on assessment centers. Meta-analytic studies have demonstrated that validities corrected for small sample size and other sources of error are about .4 (Hunter & Hunter, 1984; Schmidt et. al., 1987). While most of the validity coefficients in this study were not significant, the uncorrected coefficients between the assessment total score and the 1, 2, and 3 year performance evaluations and the supervisory overall rating were all similar to that magnitude (i.e., .31 - .40) and the coefficients corrected for the restriction of range were of course somewhat higher. Within this context, it is highly probable that all of these validity coefficients would be significantly different from zero and of a magnitude of approximately .40 or higher if a larger N had been used. In fact, given the acceptance of meta-analysis and validity generalization as sound methodological procedures, it may be that further criterion-related validity studies will only be necessary for political reasons and for evaluating new assessment center procedures or applications.

It is also interesting to point out that these findings demonstrated that the assessment center measured the interpersonal component of performance most strongly. This was evidenced by relatively strong relationships between people related predictors of criteria. Further, the performance appraisal forms were developed in 1984 and not in conjunction with the assessment center. Also, the assessors were given extensive training whereas the Fire District Chiefs/Captains who annually rated the Lieutenants were not. Nonetheless, they were consistent with their ratings on an individual over time.

## References

- Byham, W.C. (1983). Review of legal cases and opinions dealing with assessment centers and content validity. Development Dimensions International, 4, 1-43.
- Joiner, D.A. (1984). Assessment centers in the public sector: a practical approach. Public Personnel Management Journal, (1984), 13 (4), 435-449.
- Fitzgerald, L.F. The incidence and utilization of assessment centers in state and local governments. Washington, D.C.: International Personnel Management Association, 1980.
- Fitzgerald, L.F. & Quaintance, M.K. (1982). Survey of assessment center use in state and local government. Journal of Assessment Center Technology, 5 (1), 9-22.
- Hunter, J.F., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.
- Mendoza, R.H., Jr. & Craig, L. (1983). An integrated selection system for entry-level criminal justice personnel featuring the assessment center approach. Journal of Assessment Center Technology, 6 (1), 1-8.
- O'Leary, L., Bergeson, D. & Fabyan, D. (1987). Building credibility in an innovative promotional process. International Association of Fire Chiefs, 9-15.
- Schmidt, F., N., Gooding, R.Z., Noe, R.A. & Kirsh, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.
- Thornton, G.C. III and Byham, W.C. (1982). Assessment centers and managerial performance. New York: Academic Press.
- Yeager, S.J. (1986). Use of assessment centers by metropolitan fire departments in north america. Public Personnel Management. 15 (1), 51-64.



Table 1

Variable Listing with N, Means, and Standard Deviations (SD)

<u>V#</u>	<u>VARIABLE</u>	<u>N</u>	<u>MEAN</u>	<u>SD</u>
1.	Assessment Center Total Score	18, 37	32.33, 29.62	3.03, 4.90
2.	Analysis Dimension	18, 37	3.72, 3.54	.75, .73
3.	Dealing With People Dimension	18, 37	4.11, 4.0	.58, .71
4.	Decisiveness Dimen- sion	18, 37	4.61, 4.35	.50, .75
5.	Developing Subor- dinates Dimension	18, 37	3.83, 3.18	.92, 1.29
6.	Judgement Dimension	18, 37	3.78, 3.27	.55, .90
7.	Leadership Dimen- sion	18, 37	4.06, 3.54	.72, .87
8.	Planning & Organi- zing Dimension	18, 37	4.00, 3.59	.91, 1.01
9.	Verbal Communica- tion Dimension	18, 37	4.22, 3.95	.81, .85
10.	Performance Evalua- tion (4 mo)	17*	3.65	.93
11.	Performance Evalua- tion (1 yr)	16*	4.19	.54
12.	Performance Evalua- tion (2 yrs)	17*	4.18	.53
13.	Performance Evalua- tion (3 yrs)	16*	4.25	.45
14.	Analysis Dimension	17*	3.88	.78
15.	Dealing With People Dimension	17*	3.59	.94
16.	Decisiveness Dimen- sion	17*	4.06	.83
17.	Developing Subordin- ates Dimension	17*	3.71	.92
18.	Judgement Dimension	17*	3.76	.83
19.	Leadership Dimension	17*	4.06	.75
20.	Planning and Organi- zing Dimension	17*	4.18	.73
21.	Verbal Communication Dimension	17*	3.76	.97
22.	Overall Evaluation	17*	4.12	.78
23.**	Citizen Complaint	17, 36	30.00, 27.42	6.01, 5.55
24.**	Leaderless Group Discussion	17, 36	29.53, 26.06	7.89, 7.35
25.**	Coaching & Counsel- ing	17, 36	33.06, 28.22	4.07, 5.64

Continued. . .

Table 1

Variable Listing with N, Means, and Standard Deviations (SD)

<u>V#</u>	<u>VARIABLE</u>	<u>N</u>	<u>MEAN</u>	<u>SD</u>
26.	Coordinates/directs fire scenes	17*	4.1765	0.8090
27.	Evaluates subordinate performance	17*	3.8824	0.7812
28.	Promotes training/ development	17*	3.8824	0.7812
29.	Resolves disciplin- ary problems	17*	3.6471	0.7019
30.	Resolves complaints	17*	3.5294	0.7998
31.	Exchanges information	17*	4.2353	0.7524
32.	Prepares reports/ paperwork	17*	4.1765	0.8090
33.	Accepts responsibility for decisions/actions	17*	4.1765	1.0146
34.	Encourages teamwork	17*	4.2353	0.7524

\* Missing Data

\*\* Exercises

Table 2

Relationship Among Exercises And Between  
Exercises And The Assessment Center Total Score

<u>VAR.</u>	<u>Variable Label</u>	<u>N</u>	<u>MEAN</u>	<u>SD</u>
V1	Assess Ctr. Total Score	37	29.6216	4.9010
V23	Citizen Complaint	36	27.4167	5.5517
V24	Leaderless Gp. Dis.	36	26.6667	7.3485
V25	Coaching & Counseling	36	29.2222	5.6372

Simple Correlation Matrix

	<u>V1</u>	<u>V23</u>	<u>V24</u>
V23	r .5548*		
V24	r .6793*	.5722*	
V25	r .8965*	.5100*	.6991*

\* Significant at or below p = .05 with 35 df

Table 3

Relationship Among Dimensions

<u>VAR.</u>	<u>Variable Label</u>	<u>N</u>	<u>MEAN</u>	<u>SD</u>
V2	Analysis	37	3.5405	0.7301
V3	Dealing with People	37	4.0000	0.7071
V4	Decisiveness	37	4.3514	0.7534
V5	Dev. of Subordinates	37	3.1892	1.2875
V6	Judgement	37	3.2703	0.9021
V7	Leadership	37	3.5405	0.8691
V8	Planning & Organizing	37	3.5946	1.0127
V9	Verbal Communication	37	3.9459	0.8481

	<u>V2</u>	<u>V3</u>	<u>V4</u>	<u>V5</u>	<u>V6</u>	<u>V7</u>	<u>V8</u>
V3	r .5381*						
V4	r .4532*	.4172*					
V5	r .1541	.2441	.2732				
V6	r .4890*	.5661*	.5512*	.4331*			
V7	r .2272	.2260	.5928*	.4274*	.4462*		
V8	r .3046	.1552	.5196*	.4653*	.3665*	.5400*	
V9	r .0934	.2316	.2479	.0351	.3101	.2192	.2202

\* Significant at or below p = .05 with 35 df

Table 4

Relationship Between Dimensions And  
Assessment Center Total Score

---

	<u>V1</u>
V2 r	0.5789*
V3 r	0.5771*
V4 r	0.7036*
V5 r	0.6896*
V6 r	0.7777*
V7 r	0.7015*
V8 r	0.6902*
V9 r	0.3759*

---

\* Significant at or below  $p = .05$  with 35 df

Table 6

Relationship Between Supervisory Ratings on Skill Dimensions ('89)  
And  
Supervisory Ratings on Performance Appraisal Factors ('89)

---

<u>VAR.</u>	<u>Variable Label</u>
V15	Dealing With People
V17	Developing Subordinates
V20	Planning and Organizing
V32	Prepares Reports/Papers

---

	<u>V15</u>	<u>V17</u>	<u>V20</u>
V32	0.1016	0.0741	0.7933

---

\* Significant at or below  $p = .05$  with 15 df

Table 5

Relationship Between Assessment Center Ratings (1986)  
And  
Supervisory Ratings on Assessment Center Dimensions (1989)

1986					1989				
Var 2	Analysis				Var 14	Analysis			
Var 3	Dealing with People				Var 15	Dealing with People			
Var 4	Decisiveness				Var 16	Decisiveness			
Var 5	Developing Subordinates				Var 17	Developing Subordinates			
Var 6	Judgment				Var 18	Judgment			
Var 7	Leadership				Var 19	Leadership			
Var 8	Planning & Organizing				Var 20	Planning and Organizing			
Var 9	Verbal Communication				Var 21	Verbal Communication			
1986									
	<u>V2</u>	<u>V3</u>	<u>V4</u>	<u>V5</u>	<u>V6</u>	<u>V7</u>	<u>V8</u>	<u>V9</u>	
	V14	0.2689	0.4312				0.2566	0.4302	
	V15	0.2081	0.4238	0.4033				0.2912	
	V16		0.2370	0.0613			0.3232	0.4332	
1989	V17		0.5195	0.2601	0.2943			0.2597	
	V18		0.5599			0.1416		0.5372	
	V19	0.3595	0.5408	0.3975	0.2793		0.0000	0.4792	
	V20		0.3788			0.2600	0.2430	0.1837	0.4437
	V21		0.3725		0.2909	0.2359		0.2066	0.2279

\* Significant at or below  $p = .05$  with 15 df

Table 7

Relationship Between Assessment  
Center Total Score and 1, 2, & 3  
Year Performance Evaluations

	<u>V15</u> <u>1yr</u>	<u>V12</u> <u>2yr</u>	<u>V13</u> <u>3yr</u>
Uncorrected*	.35 .175	.36 .155	.40 .118
Corrected	.52	.53	.58

\* Probability due to chance with 14 df for the 1yr and 3yr correlations and 15df for the 2yr correlation appears in parenthesis.

Table 8

Relationship Between Assessment Center Total Score And  
Supervisory Ratings On Skill Dimension And Overall Performance  
(Fifth Set of Performance Data)

	<u>ANAYL</u>	<u>DWP</u>	<u>DEC</u>	<u>DOS</u>	<u>JUDG</u>	<u>LD</u>	<u>PO</u>	<u>VERCO</u>	<u>OVAL</u>
Uncorrected*	.298 (.245)	.357 (.159)	.222 (.392)	.416 (.097)	.223 (.388)	.513* (.035)	.234 (.365)	.459 (.063)	.316 (.216)
Corrected	.438	.526	.343	.594	.347	.694	.362	.640	.474

\* Probability due to chance with 1 df appears in parenthesis APPENDIX 5



TABLE 9

Relationship Between 1986 Assessment Center Total Score  
And  
1989 Supervisory Ratings on Performance Appraisal Factors

		V1 Assessment Center Total Score					V30 Resolves Complaints				
		V26 Coordinates Directs Fire Scene					V31 Exchanges Information				
		V27 Evaluates Subordinate Performance					V32 Prepares Reports/Paperwork				
		V28 Promotes Training/Development					V33 Accepts Responsibility For Actions				
		V29 Resolves Disciplinary Problems					V34 Encourages Teamwork				
		V26	V27	V28	V29	V30	V31	V32	V33	V34	
V1	Uncort.	.186 (.474)	.528 * (.029)	.528* (.029)	.653* (.004)	.389 (.127)	.469 * (.042)	.235 (.363)	.365 (.149)	.310 (.225)	
	Corrected	.291	.713	.713	.812	.564	.653	.366	.536	.475	

## Task Statements Lieutenant

1. Supervises the efforts of his/her firefighters at the station house to insure smooth operations of equipment and proper maintenance of the house and grounds, through verbal and written means.
2. Records all daily events, by writing in log book, in order to keep track of maintenance requests, number of runs, visits to station house, personnel actions, etc.
3. Delegates small firefighting problems to subordinates, through verbal commands, in order to instill confidence in the firefighter and to expand his skills.
4. Monitors job performance both at the station and the emergency scene and provides feedback and an annual Performance Evaluation by observing, recording and systematically informing his/her employees of their level of performance. This is done in order to motivate the employee and justify his/her level of merit increase.
5. Responds to emergency situations within his/her stations's area by attending to dispatcher calls, directing the firefighters under his/her command, physically riding equipment to the scene of the emergency, and calling for backup when appropriate in order to minimize the destruction and danger to life and property.
6. Directs the efforts of his/her firefighters on the scene of an emergency by evaluating the situation and verbally directing them in order to minimize the danger and destruction to human life and property (i.e. suppression and rescue).
7. Conducts training programs on technical problems as often as possible, practically from 50% to 80% of the days scheduled in order to maximize the effectiveness of all fire rescue personnel in performing the required tasks at the fire scene.
8. Confers with off-going shift lieutenants on a daily basis and reviews the log to determine shift, mechanical and other work related problems, as well as new procedures as they become available.
9. Supervises daily vehicle and equipment checks by observation and questioning the staff.
10. Counsels employees and records any personal or work related problems. Initiates disciplinary action when necessary.
11. Communicates by informing and making recommendations to the District Chief (on a verbal or written basis) on improving the overall operations of the station. (For example, recommending additional manpower to meet manning requirement and seasonal needs).

## APPENDIX 2

## Competency List For Lieutenant

1. Knowledge of Fire Suppression
  - A. Fire behavior
  - B. Building construction
  - C. Recognizing hazardous materials
  - D. Awareness of resources and capacities
  - E. Knowledge of policies, procedures, and bargaining unit
2. Knowledge of Fire Rescue
  - A. Knowledge of E.M.T.1 functions

Skill Dimensions

1. 3. Judgement-The ability to consider alternate courses of action and make decisions based upon sound logic.
2. 4. Decisiveness-Willing to make decisions and commit oneself.
3. 5. Problem Analysis-Ability to gather information and recognize inconsistencies when preparing for a decision.
4. 6. Leadership-Ability to influence individuals in a one-on-one situation or in a group setting in order to accomplish tasks.
5. 7. Development of Subordinates-Ability to take specific actions necessary in assisting the skill development of a subordinate.
6. 8. Planning & Organizing-Ability to effectively use the Resources available to accomplish tasks.
7. 9. Verbal Presentation Skills-Ability to communicate ideas verbally in a concise, complete and effective manner in both spontaneous situations and planned presentations.
8. 10. Dealing with People-Ability to consider the needs of others when accomplishing work objectives.

**USE OF NON-TRADITIONAL TRAINING AND EXPERIENCE RATINGS**

**BY THE**

**STATE OF NEW JERSEY**

**Presenters**

**Barbara Kervi**

**Susan K. Christopher**

**June, 1990  
IPMAAC Annual Conference  
San Diego, California**

## **Summary**

This paper describes the experiences of the New Jersey Department of Personnel as they shift from the more traditional training and experience ratings to the non-traditional (behavioral consistency) training and experience examinations. Several examples are described in the paper. These examples include information about the job(s), examination questions, rating procedures, and acceptance and success of the new procedures.

For a period of one year, the New Jersey Department of Personnel experimented with a variety of non-traditional training and experience ratings, gearing each test to the particular position(s) for which recruitment was conducted. Appointing authorities and candidates were satisfied and appeals from the candidates were drastically reduced.

Various approaches to non-traditional training and experience ratings are described in this paper, along with reliability information and other descriptive statistics for these approaches.

## Introduction

This paper describes the experiences of the New Jersey Department of Personnel as they shift from the more traditional training and experience ratings to the non-traditional (behavioral consistency) training and experience examinations.

For a period of approximately one year, the New Jersey Department of Personnel experimented with a variety of non-traditional training and experience ratings, gearing each test to the particular position(s) for which recruitment was conducted. Included in this paper are descriptions and examples of the various approaches to non-traditional training and experience ratings, reliability information and other descriptive statistics and a discussion of the acceptability of the new type of training and experience ratings.

Prior to February, 1989, the New Jersey Department of Personnel typically used the traditional method for training and experience ratings. In this method, points are assigned to applicants based on such factors as the number of years of training and experience, and the relevance of the training and experience to the job being sought.

In an effort to find a method of testing that would be both economical and expeditious to: reduce the number of provisional appointments; issue lists in the most cost efficient and expedient manner possible; and provide the candidates with tests that were both reliable and valid, the State Department of Personnel investigated the use of non-traditional methods for evaluating education and experience. During the next year, non-traditional training and experience examinations were developed for 40 titles, involving 79 recruitments.

### Examples

A number of different types of non-traditional training and experience examinations were used. Self-rating training and experience examinations were used when written communications was not a critical factor for the position or when the best method of collecting information about the applicant pool was a checklist, or a self-rating.

A second approach employed the behavioral consistency method; which asks applicants to describe their experiences or training and then compares the responses to a rating scale. This approach was used when the position required more complex experience or written communication skills, or when the applicant pool was predicted to be more sophisticated.



A third method combined the self-rating and the behavioral consistency method; and a fourth, included a self-rating, behavioral consistency questions, and an interest questionnaire.

Within these four methods, style and presentation of the questions were varied. Each time a non-traditional examination was administered, the problems of the administration and/or problems with the scoring of the examinations were evaluated. Subsequent examinations were then developed using the information obtained from these evaluations.

It should be noted that the applicant pool was not knowledgeable of the non-traditional training and experience examinations and the instructions for these examinations needed to be more explicit than they might have had to be with a more knowledgeable applicant pool.

#### Example One

The first non-traditional training and experience examination used by the State of New Jersey, was an examination for District Office Manager I and II and County Supervisor.

These three positions are located within the Division of Youth and Family Services, Department of Human Services and are responsible for the management of social service programs in a

district (District Office Manager I and II) or a county (County Supervisor). The difference between the two levels of the District Office Manager I and II is that the "I" manages a larger office.

Based on job analysis data, the examination dimensions for the District Office Manager I and II were, briefly:

Supervision

Policy/Program Development

Analytical Skills

Community Relations

Two additional areas were required for the County Supervisor positions:

Program Management

Budget Development

An examination was then developed for all three positions. Part A of the examination was a 20-item self-rating, with 6 items gathering information about the applicant's supervisory experience. Items 7 through 20 gathered information about the individuals analytical experience, policy/program development experience, and community relations experience.

The second part of the test was a behavioral consistency training and experience questionnaire consisting of two questions. One question asked the applicants to describe their experience in policy and procedure development and the second question asked the applicants to describe their experience with community organizations.

A third part was developed for only the County Supervisor position. This portion was, again, a behavioral consistency training and experience questionnaire. The questions asked the applicants to describe their experience and training in management, program development and budget development.

Appendix A includes this examination and the scoring criteria used. Also included in Appendix A is a copy of an enclosure which was sent to the applicant with the examination notice. The examination was administered in an "examination center."

#### Results - Example One

After the examination was administered, two raters scored the examination. The raters were a personnel analyst within the State Department of Personnel and the consultant. While it is considered good practice to use subject matter experts to rate training and experience questionnaires, it is not always feasible to do so.

The reliability of the District Office Manager I examination was .99 (N = 31,  $p=0.0000$ ) and the reliability of the County Supervisor was also .99 (N= 50,  $p=0.0000$ ).

The correlations between the parts of the examination, for District Office Manager 1, using the consultant ratings are provided below:

	Part A (7-20)	Part B
Part A (1-6)	.3559 ( $p=.049$ )	.5559 ( $p=.001$ )
Part A (7-20)		.5705 ( $p=.001$ )

While the above statistical information is interesting, one should be cautious in drawing the conclusion that the objective self-rating would be sufficient for prediction purposes.

It is also interesting to note, that with very little exposure to this methodology the personnel analyst was able to provide, independently, consistent ratings for these examinations.

#### **Example Two**

A second series of examinations were developed for Therapy Program Assistant titles. Six examinations were developed using similar format and questions for these titles.

Experience with other examinations had demonstrated the need to simplify instructions to the candidate and to provide an easy response format. The format for these examinations included a self-rating portion, involving a task checklist, an "interest inventory," and behavioral consistency questions for the senior level therapy assistants. An applicant's effectiveness in written communications was also rated for the senior level therapy assistants. Examples of the questions are provided in Appendix B.

#### **Results - Example Two**

These examinations were rated by one personnel analyst from the State Department of Personnel. While the recommended approach for rating the behavioral consistency questions is to have two

independent ratings, that is not always possible. In this case, the therapy assistant examinations contained only questions which could be objectively scored. The senior level examinations, however, did contain two questions which were rated according to an established rating scale and a rating for written communications skills.

Estimates of reliability had been obtained, however, on other examinations and it appeared from these administrations that rating scales were being used appropriately.

Of most concern for these examinations was the impact of the interest inventory. The interest inventory was a 10-item subtest which attempted to determine if the applicant was interested in working in the type of job for which the examination was given.

For the therapy assistant titles, the correlation between the score on the "experience" section of the examination and the "interest inventory" was .4332 ( $N = 32$ ,  $p = 0.013$ ). This suggests that the more relevant experience or training that the applicants had, the more they were interested in working in that environment.

This relationship did not appear for the senior therapy assistants. For three different titles, the correlations between the "experience" sections of the examination and the "interest



inventory" section were not significant. The correlations and probability levels are provided in Appendix C.

### Acceptance of Method

In the majority of cases this method of testing has been well received by the appointing authorities, the candidates, and the New Jersey Department of Personnel.

This acceptance is supported by the drastically reduced rate of appeals. Of the 1,729 questionnaires scored, there have been only 27 appeals from the candidates. Actually, since one appeal was submitted by a group of 20 persons, there have been only 8 appeals. The rate of appeal for the non-traditional training and experience examinations is less than one-tenth of a percent. Of these 27 appeals, only one candidate has carried the process beyond the first level of appeal.

Appointing authorities have also supported the use of this new methodology; raising questions in the ranking of candidates for only one candidate out of the 1,729!

The low appeal rate is particularly notable for the District Office Manager examination which was discussed earlier. In the past a multiple choice examination had been used, resulting in over half of the candidates filing appeals. Only 3 candidates

out of 146 candidates appealed the non-traditional training and experience examination.

It appears that the majority of appeals are due to the candidates' lack of familiarity with the non-traditional examination. In most cases, this was the first time they had been exposed to this type of testing.

Based on the results so far, the Department will be using this form of testing in situations where a rating of education and experience is the appropriate method of examination and the traditional method would be inappropriate. Once the candidates become more familiar with this type of testing, it is anticipated that the process will flow even more smoothly.

Copies of the appendices are available upon request to:

Barbara A. Kervi  
44 So. Clinton Ave  
CN310, 5th Floor  
Trenton NJ 08625

## CAN BIODATA PREDICT PERFORMANCE?

Herbert George Baker, PhD  
Laura E. Swirski  
Navy Personnel Research and Development Center

Somchai Dhammanungune, PhD  
Morris S. Spier, PhD  
United States International University

### ABSTRACT

A biographical information questionnaire was administered to a sample of Navy fire control technicians, who were subsequently administered an extensive hands-on, or job sample test of technical proficiency. This presentation shows samples of content from both tests, and discuss the ability of the biodata instrument to serve as a predictor of performance in this technical occupation.

With increasing pressure on traditional selection and classification testing because of equal opportunity and adverse impact, biographical information, or biodata, is being evaluated as a supplement or even a replacement for traditional tests. Also, responses to biographical questionnaires are verifiable, which may lower the propensity to cheat on entrance measures which are biographically based.

Biodata has a respectable history in the attempt to predict tenure, or survival on the job (Schuh, 1967; Asher, Kanfer, Crosby & Brandt, 1988). Major research along the same lines is even now in progress, particularly in the Department of Defense (Trent, Quenette and Pass, 1989). The prediction of performance through the use of biodata, however, is less researched and certainly less substantiated. This study is a part of the needed research in performance prediction, specifically within technical occupations.

### The Navy Job Performance Measurement Project

The Armed Services, in cooperation with the Department of Defense (DoD), are investigating the feasibility of directly linking military enlistment standards with on-the-job performance. Accurate screening and job placement are needed to ensure training success, retention of skilled personnel, and mission performance.

Linking enlistment standards and job performance requires development and evaluation of performance measures and the comparison of performance on these measures with predictors used in selection and classification. However, excessive cost and adverse impact on operational organizations of hands-on tests would prohibit DoD-wide performance testing on the scale required for validity research. Therefore, a research strategy was evolved to (1) develop a hands-on test as a high-fidelity "benchmark" against which various surrogate measures could be compared, and (2) develop less expensive, easier-to-administer surrogate measures that could substitute for the hands-on measures. Research is currently focused on more than 30 occupational specialties across the four services.

As a part of its contribution to the Joint-Service Project, the Navy (Laabs & Berry, 1987) is developing performance measures for a number of occupational specialties, or ratings, among which are the: (1) electronics technicians (ET), (2) fire controlman (FC), and (3) gas turbine systems technician-mechanical (GSM). It is in the context of the search for surrogate measures that the biodata questionnaire was constructed for use in testing FC and GSM personnel.

A biographical information questionnaire was administered to a sample of Navy fire control personnel. These technicians, who operate the data and radar elements of the MK 86 gun fire control system, were subsequently administered an extensive hands-on, or job sample test of technical proficiency.

#### Critical Task Selection

A job analysis was conducted to identify critical tasks performed by first-termers, i.e., persons in their initial period of service, generally from 1 to 6 years. "Critical" means those tasks that: (1) are performed by a sizeable number of incumbents, (2) are important to mission success, (3) have at least moderate variance in performance, (4) are representative of the job domain to the greatest extent possible, and (5) are subsumed in the first-termers job across varied duty assignments. In addition, the tasks were restricted to those involving technical proficiency, which can be accomplished on an individual basis. The result was a set of job tasks comprising the technical domain of the first-term FC. From this task list, a subset of tasks, certified by subject matter experts (SME) constituting a comprehensive, yet safe and manageable testing package, was developed.

### Hands-on Test Development

A hands-on test item was developed for each task contained in the final list of critical tasks. Each item is composed of observable, behavioral steps towards its completion. The behavioral steps for proper execution of the items are specified by SMEs, guided by manuals, written procedures, Navy policies, and safety precautions. Each step must concur with policy, and performance on each step must be readily observable. The array of hands-on test items for the FC is summarized in Table 1. (Note: The FC rating is split between data and radar subspecialties; therefore, there are actually two tests for the FC rating. Score sheets were developed for each test item, and contained a set of dichotomously scored elements corresponding to steps done correctly or incorrectly or to characteristics of task products that are acceptable or unacceptable).

TABLE 1. Hands-on Test Items

#### Radar Items

1. Perform Daily System Operability Test
2. Perform grid mode operations
3. Execute and analyze results of maintenance programs (radar)
4. Troubleshoot anti-aircraft radar equipment
5. Troubleshoot surface radar equipment
6. Adjust/align surface radar AN/SPQ-9
7. Adjust/align anti-aircraft radar AN/SPG-60

#### Data Items

1. Execute and analyze diagnostic programs
2. Troubleshoot digital computer
3. Perform Daily System Operability Test
4. Execute and analyze results of maintenance programs (data)
5. Troubleshoot gunfire control data equipment
6. Adjust/align gunfire control data equipment
7. Perform grid mode operations

### Biodata Test Development

A set of 35 factors were derived from the review of related literature. A biodata test which composed of 124 items was developed. This test was designed to elicit responses about background experiences, including high school education and activities, employment experience, military experience, and personal activities.

### Biodata Test Administration

Subjects for the biodata testing were those who were being administered the hands-on tests. All subjects were drawn from active duty personnel who were first-termers in the rating (N = 126). Subjects were logged in, given the test booklet and answer sheet, and instructed to begin. There was no time limit. Upon completion of the test, the test administrator collected the answer booklet and answer sheet. Answer sheets were scored using templates.

### Data Analysis

Data in each item in the Personal Activities Inventory was analysed by employing descriptive statistics to obtain: 1) Mean 2) Mode 3) Standard deviation 4) Percentage of choice response. Job performance data was divided into two categories according to the kinds of job performances, namely, radar operation and data processing. The raw scores of the seven sub-tasks were added up to obtain a total score for each Subject. The total score, then, was converted into standard score. The standard scores of job performance and the choice-response in each item of the Personal Activities Inventory were further analysed by employing Spearman-product moment correlation to identify significant items.

### Results

Two separate sets of biodata factors predicting job performance were identified from data analysis in the present study. The correlation coefficients of biodata factors predicting radar operation and data operation performances are shown in Table 2. and Table 3. consecutively.

TABLE 2. Correlation Coefficients of the Biodata Factors and Radar Operation Performance

1. Present level of happiness in life	.575	**
2. Machine maintenance work experiences	.504	**
3. High school grade in physics	.442	**
4. Electrical work experiences	.435	**
5. Self-rated quality of emotional control	.390	*
6. Level of motivation in the last job	.384	*
7. Preference of working alone	-.381	*
8. Self-rated ability to work with numbers	-.350	*
9. High school subjects of study	.337	*
10. Seriousness to solve problem	.322	*



TABLE 3. Correlation Coefficients of the Biodata Factors and Data Operation Performance

1. Preference of working alone	.444	**
2. Seriousness to finish a task	.430	**
3. Curiosity in how things work	.412	**
4. Experiences on machine improvement	.408	**
5. Grade in high school	.375	*
6. Manual dexterity	.366	*
7. Precision work experiences	.355	*
8. Seriousness to solve problem	.340	*
9. Self-rated quality of emotional control	.334	*
10. Self-rated quality of job performance	.325	*
11. Leadership experiences	.306	*

Note           \*\* p < .01  
                 \* p < .05

### Discussion

The first four biodata factors, which correlate to radar operation performance at a high level of significance ( $p < .01$ ), are: 1) present level of happiness in life, 2) machine maintenance work experiences, 3) high school grade in physics, and 4) electrical work experiences. The first four biodata factors, which correlate to data operation performance at a high level of significance ( $p < .01$ ), are: 1) preference of working alone, 2) seriousness to finish a task, 3) curiosity in how things work, and 4) experience on machine improvement. The only two common factors in Table 1. and Table 2. are: 1) self-rated quality of emotional control and 2) preference of working alone. However, the latter common factor correlates to the two kinds of performances in opposite directions, namely, the radar operation Subjects prefer to work with others while the data operation Subjects prefer to work alone. The findings in Table 1. and Table 2. show that the biodata factors which correlate to radar operation performance are different from the biodata factors which correlate to data operation performance.

### Conclusion

The specificity of task-related biographical factors in this study suggests that the biodata test can be employed as a surrogate for the hands-on tests. For the FC, a cross-validated study in different groups of subjects is required in order to confirm and establish the predictive factors. Further study is recommended to identify the underlying factor structure of the biodata test and to determine the relationship to success of other specific task performance.

## References

- Asher, J. J. (1972). The biographical items: Can it be improved? Personnel Psychology, 25, 251-269.
- Kanfer, R., Crosby, J. V., & Brandt, D. M. (1988). Investigating behavioral antecedents of turnover at three job tenure levels. Journal of Applied Psychology, 73(2), 331-335.
- Laabs, G. J., & Berry, V. M. (1987, August). The Navy Job Performance Measurement Program: Background, inception, and current status (NPRDC Tech. Rep. 87-34). San Diego: Navy Personnel Research and Development Center.
- Schuh, A. J. (1967). The predictability of employee tenure: A review of the literature. Personnel Psychology, 20, 133-152.
- Trent, T., Quenette, M. A., & Pass, J. (1989, August). An old-fashion biographical inventory. Paper presented at 97 th Annual Conference of American Psychological Association, New Orleans, LA.

# Can Biodata Predict Personality?<sup>1</sup>

Terry W. Mitchell  
MPORT Management Solutions  
San Diego, CA

*In the largest sense, biodata may include any and every aspect of a person's life history, including each of the many situations an individual has experienced and how the individual has behaved in those situations. Defined at an equally broad level, personality refers to the patterns of behavior that characterize each person's adaptation to situations in his or her life. However, although the definitions of biodata versus personality are strikingly similar, differences in emphasis on theory of measurement versus technology of prediction are equally striking. Several examples are available demonstrating the use of biodata technology to predict and/or measure personality traits. Such measures have the advantage of biodata's (1) high reliability and (2) resistance to faking. However, much of the power of biodata-predictors of human performance is based on the measurement of predictive factors that go beyond personality traits. Indeed, biodata capture predictive factors that are more a part of the history of the situation than they are a part of the person, and those factors are used to make valid predictions for individual outcomes.*

1. Let's begin with a definition: Biodata are factors manifest in a person's past or present situation. Note that this definition includes situational as well as individual factors. These factors are used to predict an individual's subsequent status on a criterion of interest. In a sense we can say that biodata are actually used to measure an individual's success potential in respect to a particular criterion measure.

2. Biodata can be selected and keyed to predict any reliable criterion (Anastasi, 1988), including:

A. job performance   B. turnover   C. research productivity   D. creativity

3. Scores on these kinds of biodata keys tend to have low correlations with psychological tests of individual differences such as intelligence and personality (Merenda & Clarke, 1959; Rush, 1953).

4. However, biodata may be specially developed and specifically keyed to predict individual difference scores on psychological tests (Owens & Henry, 1966), including:

A. aptitude   B. achievement   C. interests   D. personality

-----

1. Portions of this paper were presented in (1989) T.W. Mitchell (Chair), *Biodata vs. personality: The same or different classes of individual differences?* A Panel Discussion conducted at the Fourth Annual Conference of the Society for Industrial and Organizational Psychology, Boston.

One important conclusion is that biodata do not necessarily correlate with personality measures, but biodata scales can be made to correlate with, and thus to predict scores on personality tests.

5. In fact, constructs derived from factor analyses of biodata, factors variously labeled as life history dimensions, often appear similar to personality traits (Baehr & Williams, 1967; Morrison, Owens, Glennon, & Albright 1962):

A. favorable self-perception    B. extroversion    C. stability    D. drive

6. Also, biodata have actually been used to measure personality constructs (Daily, 1960; Hughes, 1956). Thus, biodata do not necessarily measure personality, but it is possible to specially develop biodata measures of personality constructs.

7. Another major point to reiterate here is that biodata can effectively predict individual differences that are generally considered as being distinct from personality, such as various measures of ability (Laurent, 1962; Owens & Henry, 1966; Sparks, 1965).

8. Beyond individual differences: Biodata are especially adept at capturing situational factors that predict individual success.

A. Here's a question for you. What is the most important factor influencing the outcome of a sales meeting? Is it . . .

- a. the personality of the buyer
- b. the asking price versus the competition's price
- c. the history of the account?

According to Theodore Higgins ("How good a sales negotiator are you?" Spirit, August, 1988), the correct answer is not: the personality of the buyer, and it is not the price of what is being sold, rather, it is the history of the account. This simply illustrates the importance of situational factors as determinants of individual success.

B. Another example is the quality of management (Ferguson & Hopkins, 1951). Within a geographically-dispersed company, simply knowing the plant, agency or office to which an individual will apply will substantially enhance the accuracy of predicting individual success in that company. This is largely because the quality of local management is a major factor influencing the likelihood and level of individual success.

C. We can see the same effect at the level of company differences (Brown, 1981). In the context of an inter-company placement function, simply knowing to which company an individual has applied would substantially improve the accuracy of predicting individual proficiency in type of a position. This is largely because of company differences in quality of management, training, compensation, staff support, and so forth.

The relevant biodata items might read something such as:

1. To what territory or region are you applying?
2. To what agency, office, or department are you applying?
3. To what company of this corporation are you applying?

These items represent choice behaviors of the individual, and are thus truly biodata, but the items are actually tapping situational and environmental factors that may be used to enhance the accuracy of predicting individual success.

9. At this point, a summary conclusion is that biodata can be developed to yield accurate and reliable measures of personality, but biodata are really very much more. For example, *recruiting source* is a biodata item often used to predict turnover (Gannon, 1971). The general finding is that recruits from personal sources (e.g., referrals) are less likely to turnover than are applicants from impersonal sources such as newspaper ads or employment agencies. It is highly improbable that these recruits from different sources have different expected turnover rates due to differences in their personalities. Rather, what really matters is the way in which a relationship between employer and employee is initiated, which is a matter of history, not personality.

The main point is that biodata are truly as rich and diverse as life itself, and we can take advantage of biodata to include situational determinants, going beyond individual difference factors, to enhance the accuracy of predicting individual success on a criterion. This high level of predictive accuracy is a tremendous practical advantage of biodata.

10. Some other practical advantages of biodata in contrast to personality measures are:
- A. biodata are accurate and reliable (Cascio, 1975; Mosel & Cozan, 1952);
  - B. biodata resist faking & falsification (Lautenschlager, 1985);
  - C. biodata need not be construct-mediated (Mitchell & Klimoski, 1982) (that is, we don't necessarily need to "work through" the latent traits underlying the predictive accuracy of biodata, assuming that our goal is mere predictive power);
  - D. biodata technology allows for the principle of equipollence (opposite personalities may be equally successful);

An example of the principle of equipollence in sales is that a very aggressive, hard sell style versus a low-pressure soft sell style may be equally effective. This is difficult for theories or technologies of personality to deal with, but biodata may accommodate this simply by using past success, achieved through any style, to predict future success.

A final conclusion relates to personality measurement for the future. If we choose to do so, we may benefit from the practical advantages of biodata by developing biodata measures of personality traits. This was recognized by Owens & Henry over 20 years ago. However, to do so will be to take advantage of only a small part of the predictive power of biodata technology, as it has been used for the past 80 years.

Over the many years, biodata has evolved as a powerful prediction technology. In this technology, we attempt to capture, and indeed to measure an individual's status on a criterion, *in advance of the actual occurrence of the criterion itself*.



This prediction technology is tremendously powerful, in part, because it includes aspects of the history of the situation as well as the history of the individual. Indeed, the ultimate power of biodata is in the capturing of extra-individual factors, such as in the history of the situation, and the use of those factors to enhance the predictive accuracy of *individual outcomes* such as job performance.

#### References

- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Baehr, M. E., & Williams, G. B. (1967). Underlying dimensions of personal background data and their relationship to occupational classification. *Journal of Applied Psychology*, 51, 481-490.
- Brown, S. H. (1961). Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology*, 66, 664-670.
- Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology*, 60, 767-769.
- Daily, C. A. (1960). Life history as a criterion of assessment. *Journal of Counseling Psychology*, 7, 20-23.
- Gannon, M. J. (1971). Source of referral and employee turnover. *Journal of Applied Psychology*, 55, 226-228.
- Hughes, J. L. (1956). Expressed personality needs as predictors of sales success. *Personnel Psychology*, 9, 347-357.
- Laurent, H. (1962). Early identification of management talent. *Management Record*, 24, 33-38.
- Lautenschlager, G. (1985). Within subject measures for the assessment of individual difference in faking. *Educational and Psychological Measurement*, 46, 309-316.
- Merenda, P. F., & Clarke, W. V. (1959). The predictive efficiency of temperament characteristics and personal history variables in determining success of life insurance agents. *Journal of Applied Psychology*, 43, 360-366.
- Mitchell, T. W. & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology*, 67, 411-418.
- Morrison, R. F., Owens, W. A., Glennon, J. R., & Albright, L. E. (1962). Factored life history antecedents of industrial research performance. *Journal of Applied Psychology*, 46, 281-285.
- Mosel, J. L., & Cozan, L. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology*, 36, 365-369.
- Owens, W. A., & Henry, E. R. (1966). Biographical data in industrial psychology: A review and evaluation. Greensboro, NC: The Richardson Foundation.
- Rush, C. H., Jr. (1953). A factorial study of sales criteria. *Personnel Psychology*, 6, 9-24.
- Sparks, C. P. (1965). *Using life history items to predict cognitive test scores*. Houston, TX: Author.



Recent Trends in Legal Requirements and Personnel Selection  
In the Public Sector

Gerald V. Barrett  
The University of Akron

RUNNING HEAD: Recent Trends

Paper presented at a Symposium on Personnel Selection Which Meets  
the Evolving Legal Requirements at the 1990 Annual IPMA Assessment  
Council Conference on Personnel Assessment, San Diego, CA, June  
28, 1990.

## Introduction

The last several years have been an exciting time for those who are avid followers of Supreme Court decisions. Recent Supreme Court decisions have clarified a number of issues. This includes adverse impact analysis as a basis for a prima facie case of discrimination, the role of subjective selection procedures and the use of stereotyping as evidence. In my view the recent Supreme Court decisions have built upon past cases and clarified the role of those actively engaged in personnel selection in the public sector. There has been much disagreement and controversy concerning the recent Supreme Court decisions and Congress may well pass new legislation which again will have to work its way through the courts and I fear muddy waters which the recent Supreme Court decisions have clarified.

In many ways those involved with personnel selection in the public sector have had a respite from some selection problems because of affirmative action programs and consent agreements. Because of the recent Supreme Court decisions I think the "Honeymoon" is now over. Recent Supreme Court decisions concerning both affirmative action and consent agreements will soon be changing the ground rules and I foresee a new round of litigation, particularly if Congress passes a new Civil Rights Act. The challenge to those involved in selection in the public sector will be to maintain the merit principle and at the same time reduce potential adverse impact.

### Affirmative Action Court Decisions

There have been over a dozen Supreme Court cases dealing with affirmative action issues both in the public and private sector (Boston Firefighters Union, Local 718 v. Boston Chapter, NAACP, 1983; City of Richmond v. J. A. Croson Company, 1989; Defunis v. Odegaard, 1974; Firefighters Local Union No. 1784 v. Stotts, 1984; Fullilove v. Klutznick, 1980; Furnco Construction Corporation v. Waters, 1978; Johnson v. Transportation Agency, 1987; Local 28 of Sheetmetal Workers International Association v. EEOC, 1986; Local 93 v. City of Cleveland, 1980; Martin v. Wilks, 1989; McDonald v. Santa Fe Trail Transportation Company, 1986; Mississippi University for Women v. Hogan, 1982; Regents of the University of California v. Bakke 1978; United States v. Paradise, 1987; United Steelworkers of America v. Weber, 1979; W. R. Grace v. Local Union 759, 1983; Wygant v. Jackson Board of Education 1976). We can now identify some general principles concerning affirmative action which have been made clear by the recent Supreme Court decisions.

First, an affirmative action program has to identify a specific problem of past discrimination or underrepresentation. An organization can not rely upon general concepts such as "societal discrimination" or the lack of "role models" to justify an affirmative action plan which is going to have an impact upon personnel selection or promotions. This does not mean that the employer has to admit past discrimination. The evidence which the

organization considers does not have to rise to the level required of a prima facie case under Title VII.

Second, the organization can formulate an affirmative action plan once a specific problem has been identified. The plan which is formulated should have a narrow focus and be a remedy for a specific problem identified in the organization. In particular, this means that it would be inappropriate for an organization to have an across-the-board plan which gave more weight in selection to minorities in all job classifications unless there was evidence that there was a specific problem in all job classifications.

Third, the goals and time tables of an affirmative action plan should focus upon the specific identified problem. If the problem is one in initial selection then that is what the goals and time tables should focus upon. It would probably be inappropriate to have goals and time tables which specify a specific promotion rate for minorities.

Fourth, the goals of the affirmative action plan should be to obtain but not maintain a balanced workforce. Once a stated goal is reached then the additional weight which might have been given to a minority group applicant would no longer be appropriate.

Fifth, any affirmative action program should not unnecessarily abridge the rights of nonminorities. There will now have to be more of a balance between the rights of minorities and nonminorities when an affirmative action program is implemented.

Sixth, the preferred approach to an affirmative action program appears to be some variation of the "Harvard Plan." This concept was endorsed in Bakke where gender or race is just one factor which is given consideration in the personnel decision. For example, if these test scores are relatively close then minority status may be given a positive weight in the final selection decision.

Seventh, the courts are making it increasingly clear that only they have the power to order quotas. It would be improper for an organization to establish its own quotas unless there has been some judicial order.

Eighth, the burden of persuasion is upon the challenger to show an affirmative action program is not valid. This is a clear advantage for an organization which formulates a reasonable affirmative action program.

Because of the affirmative action programs that have been instituted by some public organizations the selection specialist has been under a protective umbrella. There has been no adverse impact and the selection system was not challenged. This I think is now ending. Personnel selection systems can no longer have the freedom to operate under the umbrella of an affirmative action program which assures the organization of no adverse impact. As there has been a change in affirmative action programs there has also been a change in consent agreements.

Consent Agreements

A favorite technique to resolve employee discrimination suits in the public sector has been to enter into consent agreements with the plaintiff. These consent agreements often specify that a certain number of minorities must be selected and or promoted each time there are openings in the organization. This again protected most selection systems from scrutiny by litigious plaintiffs. As in the case of an affirmative action program there would be no evidence of adverse impact and therefore it would be extremely difficult for a plaintiff to prevail under Title VII.

The Supreme Court in *Martin v. Wilkes* (1989) changed the legal posture of consent agreements. The decision in effect repeats the well established principle of law that people can not be deprived of their legal rights in a proceeding in which they are not a party. This basically means that it is going to be much more difficult to have consent agreements which specify that a certain percentage of minorities will be hired or promoted. It is clear that the Supreme Court has an aversion to quotas (*City of Richmond v. Croson*, 1989).

The protective shields of affirmative action and consent agreements are now being removed. Personnel selection will have survive on its own. At the same time we now have more of a reproachment between professional selection literature and recent court decisions.



Convergence in the Personnel Selection Literature and  
Recent Supreme Court Decisions

After *Griggs v. Duke Power* (1971) was decided, the conventional wisdom was that it would be impossible to use tests because of the strict standards imposed by the Supreme Court. There are those who contended that only a predictive validation study using actual job performance would pass muster in the courts. They believed that even a concurrent validation study was not adequate to show that a selection procedure was valid (Barrett, Phillips, & Alexander, 1981).

The Trinity of construct, content, and criterion-related validity has been displaced. There is now a realization both by professionals in the field and the courts that there are many ways to produce evidence that a selection procedure is job related (Barrett & Kernan, 1987; Binning & Barrett, 1989; Cronbach, 1988; Landy, 1986; Messick, 1989).

Construct/Content Validation Models

The importance of constructs to validation have received more emphasis (Binning & Barrett, 1989). This is a complex topic but validation is best considered a process of hypothesis testing (Landy, 1986). This broader conception of the validation process has led to use of the term construct/content validation models.

At one point in time the plaintiffs experts argued vigorously that content validity was never appropriate. Content valid tests were not accepted as evidence a selection procedure

was job related. Now both professionals and the courts are accepting the use of a construct/content validation approach. (Frederikson, 1986; Pulakos, Borman, & Hough, 1988; Schmitt & Ostroff, 1986; Turban, Sanders, Francis & Osburn, 1989). The construct/content validation approach has been accepted and used for entry-level examinations (Guardians v. City of New York, 1980, Zamlen v. City of Cleveland, 1988).

We now have considerable evidence that the educational system does not allow us to assume that individuals have basic knowledge and skills which are required even for the most simple entry-level jobs. The use of a content valid test will continue to be an important selection tool.

The Supreme Court has consistently found it much more reasonable than the plaintiffs experts that training is a suitable criterion and measure of job performance (Washington v. Davis, 1976; Watson v. Fort Worth Bank, 1988). Training programs are usually the employees first job assignment. This is certainly a measure of job performance (Reilly & Israelski, 1988). Training performance will continue to be important in criterion-related validation studies.

#### Validity Generalization

The plaintiffs bar has been very concerned about validity generalization. Their concern focuses upon their fear that validity generalization will become the sole evidence that the organization needs to present in order to show that their

cognitive ability tests are job related (Goldstein & Patterson, 1988; Seymour, 1988). Both court decisions and professional opinion makes this outcome very unlikely (EEOC v. Atlas Box, 1989; Hartigan & Wigdor, 1989). Validity generalization does have a role in helping the test constructor develop and defend their selection instruments. It is best considered just additional evidence to support the job relevance of certain selection procedures.

A transfer study involves taking a specific selection procedure validated in one location and transferring it to a second location in a different organization (Friend v. Leidinger, 1978). This approach has been successfully defended for firefighters (Brunet v. City of Columbus, 1986).

#### Subjective Selection Procedures

At one point in time it was very common for attorneys to recommend that subjective selective procedures be used because they did not have to be validated. Employers believed the interview process was immune from the usual requirements of Title VII. There is now no question that subjective selection procedures can be tested under the disparate impact theory of discrimination (Watson v. Fort Worth Bank, 1988). The obligation of the employer is the same with subjective or objective selection procedures to show that they are job relevant if there is disparate impact.

The situation becomes even more complicated if there can be an assertion that there is a stereotype operating which may have influenced the personnel decision. The Supreme Court accepted the idea of stereotypes as evidence of gender discrimination (Price Waterhouse v. Hopkins, 1989). If stereotypes are evidence of gender or race being a factor in the employment decision then we have a mixed-motive cases. In this situation the burden of persuasion would fall on the employer to show that the selection procedure was job relevant. This sort of case presents a very difficult burden for the employer (Barrett, 1990). All subjective personnel selection procedures will have to be constructed with a great deal more care to be sure they are free from any stigma of stereotypes and are job relevant.

#### Adverse Impact Analysis

The typical first step by a plaintiffs expert is to establish that there has been adverse impact because of the use of the selection procedure. A prima facie case has been established for the plaintiff and the employer has the burden of coming forward with some evidence that the selection procedure is job relevant. This first hurdle for the plaintiff has usually been relatively easy. The plaintiffs expert will show a discrepancy between the general population and minority percentage in a job category. Alternatively, the internal work force figures will be used to show that the minority percentages are not uniform in all job categories.

The Court in *Wards Cove Packing Company v. Antonio* (1989) has changed the disparate impact analysis. A mere statistical imbalance between different job groups in an organization does not establish a prima facie case of discrimination.

What has often been called Simmons Paradox an Aggregation Fallacy, Ecological Fallacy, or Berkson's Fallacy (Alexander, Barrett, Alliger, & Carson, 1986; Bickel, Hammel, & O'Connell, 1975; Robinson, 1950) has been eliminated.

This paradox occurs when an occupational category having the lowest promotion or selection rate in an organization also has the greatest proportion of minorities. It is ironic that this phenomena can occur when the separate promotion and selection rates for each job category may actually favor the minority group.

#### Conclusion

Part of the function of the law is to build up through a series of cases precedent which can guide the organization. The idea is that the attorney can give advice to organizations as to the appropriate action based upon past precedent. We are now in the fortunate position of having past precedent which will guide us in developing selecting systems which have utility for an organization and at the same time meet prevailing legal requirements. The main issue will be to maintain merit principles and at the same time reduce any potential adverse impact. If Congress does pass the Civil Rights Act of 1990 then the process will start over again. It will be years before organizations can

develop and implement selection systems they can reasonably be sure meet the prevailing legal requirements. The happy convergence between the professional research and the Supreme Court decisions may be obliterated by Congress and will start over the long road we have walked since 1971.



## References

- Alexander, R. A., Barrett, G. V., Alliger, G. M., & Carson, K. P. (1986). Toward a general model of nonrandom sampling and the impact on population correlation: Generalizations of Berkson's Fallacy and restriction of range. British Journal of Mathematical and Statistical Psychology, 39, 90-105.
- Barrett, G. V. (1990). Personnel selection after Watson, Hopkins, Atonio, and Martin (WHAM). Forensic Reports, 3, 179-203.
- Barrett, G. V., Phillips, J. S., & Alexander R. A. (1981). concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 66, 1-6.
- Barrett, G. V., & Kernan, M. C. (1987). Performance appraisal and terminations: A review of court decisions since Brito v. Zia with implications for personnel practices. Personnel Psychology, 40, 489-503.
- Bickel, P. J., Hammel, E. E., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. Science, 187, 398-404.
- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied psychology, 74, 478-494.
- Boston Firefighters Union, Local 718 v. Boston Chapter, NAACP, Inc., 468 U.S. 1206 (1983).
- Brunett v. City of Columbus, 642 F.Supp. 1214 (S.D. Ohio 1986).

City of Richmond, v. J. A. Croson Company, 109 S. Ct. 706 (1989).

Cronbach, L. J. (1988). Five perspectives on validation argument.

In H. Wainer & H. Braun (Eds.), Test validity (pp. 3-17).

Hillsdale, NJ: Erlbaum.

Defunis v. Odegarrrd et al., 416 U.S. 312 (1973).

Equal Employment Opportunity Commission v. Atlas Paper Box

Company, 868 F.2d 1487 (6th Cir. 1989).

Firefighters Local Union No. 1784 v. Stotts, 467 U. S. 561 (1984).

Frederiksen, N. (1986). Construct validity and construct

similarity: Methods for use in test development and test

validation. Multivariate Behavioral Research, 21, 3-28.

Friend v. Leidinger, 588 F.2d 61 (1978).

Fullilove v. Klutznick, 448 U.S. 448 (1980).

Furnco Construction Co. v. Waters, 46 U.S.L.W. 4966 (1978).

Goldstein, B. L., & Patterson, P. O. (1988). Turning back the

Title VII clock: The resegregation of the American work

force through validity generalization. Journal of

Vocational Behavior, 33, 452-462.

Griggs v. Duke Power, 401 U.S. 424 (1971).

Guardians Association of New York City v. Civil Service Commission

of City of New York, 630 F.2d 79 (1980).

Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). Fairness in

employment testing: Validity generalization, minority

issues, and the General Aptitude Test Battery. Washington,

DC: National Academy Press.

Johnson v. Transportation Agency, Santa Clara County, California,  
480 U.S. 616 (1987).

Landy, F. J. (1986). Stamp collecting versus science: Validation  
as hypothesis testing. American Psychologist, 41(11), 1183-  
1192.

Local 28, Sheet Metal Workers' International Association v. EEOC,  
478 U.S. 421 (1986).

Local 93, International Association of Firefighters v. City of  
Cleveland, 478 U.S. 501 (1986).

Martin v. Wilks, 109 S.Ct. 2180 (1989).

McDonald v. Santa Fe Trail Transportation, 427 U.S. 273 (1976).

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational  
measurement (3rd ed.) (pp. 13-103). New York: Macmillan.

Mississippi University for Women v. Hogan, 458 U.S. 718 (1982).

Price Waterhouse v. Hopkins, 109 S.Ct. 1775 (1989).

Pulakos, E. D., Borman, W. C., & Hough, L. M. (1988). Test  
validation for scientific understanding: Two demonstrations  
of an approach to studying predictor-criterion linkages.  
Personnel Psychology, 41, 703-716.

Regents of the University of California v. Bakke, 438 U. S. 265  
(1978).

Reilly, R. R., & Israelski, E. W. (1988). Development and  
validation of minicourses in the telecommunication industry.  
Journal of Applied Psychology, 73, 721-726.

- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15, 351-357.
- Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. Personnel Psychology, 39, 91-108.
- Seymour, R. T. (1988). Why plaintiff's counsel challenge tests, and how they can successfully challenge the theory of "validity generalization." Journal of Vocational Behavior, 33, 331-364.
- Turban, D. B., Sanders, P. A., Francis, D. J., & Osburn, H. G. (1989). Construct equivalence as an approach to replacing validated cognitive ability selection tests. Journal of Applied Psychology, 74, 62-71.
- United States v. Paradise, 480 U.S. 149 (1987).
- United Steelworkers of America, AFL-CIO-CLC v. Weber, 443 U.S. 193 (1979).
- W. R. Grace & Co. v. Local Union 759, 461 U.S. 757 (1983).
- Wards Cove v. Atonio, 109 S.Ct. 2115 (1989).
- Washington v. Davis, 426 U.S. 229 (1976).
- Watson v. Fort Worth Bank and Trust, 487 U.S. 977 (1988).
- Wygant v. Jackson Board of Education, 467 U.S. 267 (1986).
- Zamlen v. City of Cleveland, 686 F.Supp. 631 (N.D. Ohio 1988).

Pressures to Use Unwarranted Procedures  
To Reduce Adverse Impact in Public Sector Personnel Selection

Ralph A. Alexander  
The University of Akron

## Pressures to Use Unwarranted Procedures

### To Reduce Adverse Impact in Public Sector Personnel Selection

In the public sector, at least as much as in the private sector, there is an imperative that our personnel selection procedures meet the societal demands imposed through legal constraints on our practices but at the same time meet the merit principle by making full use of the best current state-of-the-art information on the science and practice of personnel selection.

From time-to-time procedures emerge, quickly acquire a vocal cadre of advocates, and gain great currency because they seem to satisfy both sets of requirements--particularly they seem to provide, at face value, the appearance of improved fairness. A more thoughtful analysis, however, often reveals that, in reality they are neither more fair nor do they represent good science. I want to discuss two recent examples that are presently being advocated in some quarters: the first is the so-called "Golden Rule" for retaining or discarding test items; the second is test-score-banding as an alternative to top-down hiring or top-down referral based on applicant test scores.

I have two purposes in discussing these particular procedures. The most obvious, of course, is to point out the pitfalls in these procedures specifically. My second, broader



purpose is to encourage you to carefully inspect any newly emerging fad. If it seems to provide too easy an answer--as is the case with the Golden Rule--then it may indeed be too good to be true. Or, if it seems to fly in the face of sound logic--as is the case with test-score banding--then it may rest on false premises.

The Golden Rule is really nothing more than the simplest, and seemingly most straightforward of the statistical procedures for analyzing test items for apparent bias against minority group members. Briefly, the Golden Rule proposes that if the passing rate on a test item is lower for the minority group than for the non-minority group by more than some fixed amount (usually a 15 percentage point difference is used) then the item is discarded under the presumption that it is biased.

The past 10-15 years has seen a great deal of interest and published research on item-bias analysis methods. Nowhere in that literature do we find a serious apologist for using the simple algebraic difference in group item-difficulty levels as an accurate or an adequate method for detecting item bias. There are at least three very good reasons that the method is not recommended. First, proportions or percentages are notorious for having inordinately large sampling errors. With the samples and

sample sizes typically used for item analysis of selection tests, group differences in passing rates of ten percentage points in one sample and 20 percentage points in the next sample are well within expectation solely on the basis of random sampling. Thus, the decision to retain or discard an item based solely on differential group difficulties is likely to be highly unstable from one applicant group to the next. Second, there are sound statistical and empirical reasons to believe that eliminating items on this basis will have little or no effect on the overall adverse impact of the total test-score. Third, there is also growing evidence that eliminating items on this basis may substantially reduce test validity.

Two years ago the American Psychological Association meeting contained a public-policy symposium on the Golden Rule where the issues were discussed in substantially more detail than we have time for here. If you are presently using or are contemplating using the Golden Rule for test construction, I would strongly recommend that you seek out the papers that were presented at that symposium and that you reconsider the use of the Golden Rule as a basis for retaining or rejecting test items.

The idea of test-score banding is currently being promulgated as a method that is more fair to minority group

members than top-down selection. The procedure involves some variant of the following scheme: Beginning with the highest observed test score in an applicant group, a 95% confidence band of scores is found based on the standard error of measurement. Individuals are selected randomly from within this band until the band is depleted, then a new band of test scores is similarly constructed from the remaining scores and selection again proceeds at random. The only mention of this concept to appear in the literature is a brief discussion in a 1984 Public Personnel Management Journal article by Sproule who discusses banding but makes no reference to random selection within bands. No other discussion or consideration of the procedure has appeared in the literature. In spite of this, it has either been used or been advocated in a number of court cases, almost all of them public sector cases. The rationale given by those who recommend the procedure is that since tests are not perfectly reliable, test scores within  $\pm 1.96$  standard errors are not statistically significantly different from each other and should, therefore, be treated as being equal to each other.

The flaws in this procedure are numerous, both logical and practical. The fundamental error here is the error of equating "no statistically significant difference" with "equality." The

two inferences are simply not the same thing. To say that the inference of no difference is the same as the inference of equality is an error that has been pointed out in the literature on logical inference for at least 300 years. Second, how adding random error to all test scores within some band (and that is precisely what random selection within the band is doing) in any way improves either the quality or the fairness of the selection decision simply escapes me. I'd like to quote from the 1987 2nd Circuit Court of Appeals decision in the case Berkman vs. City of New York: "The facts did not warrant the remedy of random assignment of written test scores. That remedy unfairly deprives many applicants of the enhanced opportunity they derived by scoring comparatively better on the test." The judges in this case seemed to have as much trouble with the logic of adding random error as I have.

The procedure also forces a logical paradox. Consider a 100 point test that results in 10-point bands, e.g., 91-100, 81-90, etc. Under this scheme scores of 91 and 100 [a 9 point difference] would be treated as though they were equal because they are not statistically significantly different from each other. On the other hand, scores of 90 and 91 [a 1 point difference] (being in different bands) would be treated as being

different from each other in spite of the fact that they are also not significantly different from each other. Flawed logic often leads to this sort of paradoxical outcome.

Third, even if we were willing to accept the argument that at some point observed test scores (because they are fallible) should be treated as though they were not meaningfully different from each other, an extensive technical debate would then revolve around the "operational definition of statistical equality." There are a number of perfectly legitimate alternatives to the standard error of measurement definition. For example, a strong argument could be made for using a definition based on the probability that the highest and lowest test scores in the band would reverse their rank-order on repeated testing. In the case of a test score band of  $\pm 1.96$  SEM, that probability is slightly less than 10%. Is that sufficient justification for a judgment of "no meaningful difference?" I don't think so.

In the public sector, there is also, it seems to me, a problem of the governing statutes. In almost every civil service jurisdiction, the law requires that the results from selection procedures be rank-ordered. Challenges to this requirement have routinely failed. In my view, deliberately adding random error to

test scores, or treating all scores within a broad band as being equal, subverts this requirement.

There is also a serious question regarding the empirical (criterion-related) validity of test scores when banding is practiced. The relevant validity is not the correlation between observed test scores and job performance. Rather, the relevant validity is the correlation between test score plus the random error component vs. job performance. Would anyone like to guess what that will do to your validity coefficient?

In a different vein, a practical question occurs to me--does it make a difference? That is, can I expect test-score-banding to reduce adverse impact vis-a-vis top-down selection? In general, the answer is a resounding "No!" With the exception of very strange test-score distributions, test-score banding will, on average, produce adverse impact results that are virtually identical to top-down selection.

The operative phase here is "on average"--another implication of this is that there is about a 50-50 chance that test-score banding will result in less adverse impact than top-down selection. There is also about a 50-50 chance that the result will be worse!



Which brings me to my final practical concern. Consider the unsuspecting personnel manager who creates an eligibility list based on test-score banding. It occurs to him or her to check for adverse impact and finds that by chance there is a more adverse impact than under strict top-down selection. What would you do? My guess is that most people would quietly steal back to the computer and keep trying until they get the desired result.

In closing, let me reemphasize what I said earlier--until the latest splash has seen the light of serious discussion, beware!

Information Processing Approaches to  
Test Development and Construction as Evidence for Test Validity

Dennis Doverspike

University of Akron

Running Head: IP

Paper presented at a Symposium on Personnel Selection Which Meets  
the Evolving Legal Requirements at the 1990 Annual IPMA Assessment  
Council Conference on Personnel Assessment, San Diego, CA, June  
28, 1990.

Information Processing Approaches to  
Test Development and Construction as Evidence for Test Validity

The last decade was marked by many challenges to the tenets of personnel selection research. This was especially true of many of our ideas concerning the primacy of predictive studies as a validation strategy. During the last decade, we saw a rapid growth in our understanding of validation strategies and the introduction of a number of alternatives to traditional criterion-related validation including meta-analysis, validity generalization, content oriented behavioral consistency approaches (Schmitt & Ostroff, 1986) and content/construct strategies (Binning & Barrett, 1989; Borman, Rosse & Abrahams, 1980; Landy, 1986). We have also seen an increased emphasis on the importance of the g factor in testing. One promising outcome of the rise of new validation strategies, has been an increased concern with theory development. This has lead many applied psychologists to look to recent advances in experimental psychology, especially artificial intelligence research, where progress has been made in developing new models of human information processing (IP) as well as measurement techniques.

The last decade also saw technological change, with the introduction of the use of personnel computers in the personnel

field. In the area of testing, personal computers have already simplified the administration of tests and the collection of data. In addition, they also create the opportunity for the development of tests never before possible using a traditional paper-and-pencil test.

As we look forward to the year 2000, the computerized testing of IP abilities, seems to offer unlimited growth potential and represent an attractive alternative to traditional paper-and-pencil tests of general ability. However, to this point in time, computerized IP tests have yet to live up to their full potential.

For the last 15 years, researchers at the University of Akron have been working on the problem of the development of IP based tests. The research has combined field and laboratory research.

#### Definition

To start, we might ask what is an IP test. It is an ambiguous term and can refer to many different perspectives in cognitive and social psychology. In general, when we speak of an IP test we are speaking about a test which was based on an approach which viewed the human mind as an information processor that codes, stores and retrieves environmental inputs. A more specific definition would be that it is a general term for a class

of both paper and pencil and computerized tests in which the subject's recognition, retention, reaction time, capacity or other responses are measured in response to verbal or perceptual stimuli. Another way of thinking of IP approaches is that they are cognitive consistency approaches. That is, in developing items on the test we try to create items which will create a press for cognitive operations which are isomorphic to those required by the job. This may then be seen as differing from general cognitive ability, behavioral consistency approaches, or job knowledge approaches.

As mentioned previously, we have been working on various types of IP tests for the past 15 years in both lab and field studies. This approach is applicable to a variety of job types. Jobs which have been studied include secretarial, clerical, fire fighter, police officer transport driver, process control, maintenance mechanic, and radar operator.

#### Development

When developing IP tests, there are a number of approaches which can be followed. While IP approaches appear to be particularly compatible with computerized testing, it is possible to develop paper-and-pencil IP tests. For paper-and-pencil testing, two basic developmental approaches have been used. The

first is to adapt or select a test based on currently available IP tests based on the IP constructs identified through a job analysis. The second approach would be to build or construct a test based on a construct/content strategy of test development.

For computerized IP tests several options are also available. One simple option is to simply convert a paper-and-pencil test into a computerized format. A second option is to turn to the experimental psychology literature, look for successful measures, and turn these measures into computerized IP tests. The third option would be to build or construct a test based on a construct/content strategy.

Regardless of the approach followed, the first step would be a job analysis. The IP tests construction process requires a much more detailed, intensive approach than might be normally required, although with experience this need not be a torturous experience. It does mean that we can not simply go out and ask job incumbents what IP's are required on their job. Rather the job analyst deduces the IP elements from an analysis similar to one used in protocol analysis. That is incumbents are asked to think/talk aloud about performance. We call such an analysis an IP job analysis (Barrett & Maurer, 1989; Arthur, Barrett & Doverspike, 1990).



If we look at the job analysis traditionally we work at the behavioral level. The IP approach goes beyond this to the IP level. With the macro IP approach we are basically looking at decision points or what types of cognitive operations are being carried out in a general sense. Examples might be retrieving names from memory or a particular method of scanning a map. At the micro IP level we are attempting to extract even more information, for example how many bits of information have to be held in memory and for how long.

The second step is the identification of relevant constructs/content. This follows naturally from the job analysis. Look to see what are the critical IP tasks which can be measured. We may have many IP tasks. However, for some there will not be individual differences. For others, we may not be capable of measuring them or we cannot measure them in a reasonable fashion. This allow us to limit the set of IP elements to be measured. We then need to link IP elements to job behaviors.

The third step is the actual development of tests. Here, we can follow standard test development procedures including the development of appropriate instructions, decision on types of stimuli, number of stimuli, types of responses, numbers of

responses, and types of measures (for example reaction time or number correct).

The fourth step is pilot testing. This may include experimental studies manipulating characteristics of the tests as well as traditional item analysis.

The fifth step is the development of the final version of the test. The sixth step is a continuation of the second step. This step involves documenting the links from job behaviors to IPs to test questions.

The seventh step involves studies of the validity of the tests as predictors of simulator performance. One of the unique aspects of our program has been the use of simulator performance as a criterion. Traditionally, we have lionized performance appraisal ratings as a criterion despite their well known weaknesses. Our argument is that a well constructed high fidelity simulation can serve as a more effective criterion than those often employed in criterion related study. The development of the simulator is also linked to the IP job analysis.

#### Conclusion

Computerized IP testing will continue to grow in importance as we move to the year 2000. While a number of problems still need to be solved, both computerized and IP testing of abilities

is already a viable alternative to traditional testing. Unlike g measures, IP approaches can link testing and training, and serve as a foundation to both

## References

- Arthur, W., Jr. (1987). The validity of information processing measures predicting accidents in simulated and real world contexts. Unpublished doctoral dissertation, The University of Akron, Department of Psychology, Akron, OH.
- Barrett, G. V., & Maurer, T. J. (1989). The job analysis-predictor development process: Beyond g and generic constructs in employment aptitude testing. Unpublished manuscript, The University of Akron, Department of Psychology.
- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, 74, 478-494.
- Borman, W. C., Rosse, R. L., & Abrahams, N. M. (1980). An empirical construct validity approach to studying predictor-job performance links. Journal of Applied Psychology, 65, 662-671.
- Landy, F. J. (1986). Stamp collecting versus science. American Psychologist, 41, 1183-1192.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the  
"behavioral consistency" approach: Selection test  
development based on a content-oriented strategy. Personnel  
Psychology, 39, 91-108.

## **Candidate Preparation for Assessment Centers**

**Janine P. DuMontelle**  
**Senior Personnel Analyst**  
**University of California at Irvine**

There has been increasing discussion among personnel professionals regarding the role of candidate preparation for Assessment Centers. There are some obvious benefits or positives, such as reduced candidate anxiety and starting all candidates off with the same information regarding the process. The idea being, to make everyone as comfortable as possible, having a reasonable amount of information, vis a vis, what can be expected during testing. This can normally be accomplished in an orientation session.

There can also be some negatives associated with candidate preparation. These vary depending on the type of preparation received by the candidate, but, tend to include, gameplaying, misunderstanding of information and over emphasis on guidelines from training.

There is some agreement among professionals that certain information should be provided to candidates. Candidates should know:

- \* the dimensions being measured
- \* the exercises
- \* the rules
- \* the protocol

On the other extreme there is also some training of candidates which involves strategies such as the "water cooler" technique : hang around the water cooler listening to candidate that have completed testing and pick up as much information regarding the test as possible.

There seems to be a continuum. Most professionals are pretty clear that giving the orientation information is good and should occur, and are also clear that the "water cooler" advice is compromising the integrity of the test and should not occur. There is however a gray area in between.

Some of the issues that fall in the "gray area" are:

- \* how candidates are graded
- \* what dimensions are in each exercise
- \* who the assessors are in advance
- \* practicing exercises



The following are some observations and concerns regarding preparation that goes beyond orientation:

- \* Gameplaying - Saying what they think assessor want to hear
- \* Misunderstanding information - No matter how carefully information is presented it is sometimes misunderstood. Sometimes the information is misprocessed - people hear what they want to hear.
- \* Misused information - The information was clear when presented , but , there is trouble applying it. Practicing encourages candidates to do things the way they remember them being demonstrated , therefore , the training can interfere with their performance , in that they do not rely on their own experience because they believe there is a technique. This not something trainers reinforce - candidates doubt themselves.

## CASESTUDY

### SGT X

This candidate had attended a preparation session in which exercises were practiced. This training session was sponsored by the police department that he worked for and was made available to all candidates.

During the administration of the assessment center , the candidate realized that the instructions for an exercise was different that those presented during the training session. Instead of asking a question, he assumed that the training information was more accurate that the Assessment Center instructions. **RESULT:** The candidate failed the hurdle and was eliminated from the process.

In reviewing with the candidate in a feedback session it was revealed that he felt he should take the approach demonstrated in the training, after all, the Chief brought in the trainer, surely that was the approach that the Chief preferred.

**IMPACT:** If the candidate had not been given prior instruction he would have had no reason to doubt the Assessment Center instructions. In his words the preparation " caused me to doubt myself, doubt my interpretation of the instructions and my course of action." In this instance the preparation and training interfered with the candidate's ability to perform.

It has been my experience that training, that goes beyond orientation , puts the Assessment Center developer/administrator in the position of having to defend either the Assessment Center design, the training or both. Candidates feel that they have two equally valid methods for administration and since the training comes before the Assessment Center , candidates tend to latch on to the training perspective.

I will put the question to my colleagues : Are we creating more problems than we are solving?

# **INNOVATIVE METHODS FOR CUT SCORE DETERMINATION AND INJURY PREVENTION THAT IMPROVE THE BOTTOM LINE (\$)**

***Deborah L. Gebhardt, Ph.D.  
Human Performance Systems, Inc.***

## **Introduction**

This presentation will encompass two studies that use biomechanical modeling and analysis to determine cutoff scores for physical performance tests and to provide workplace alterations that reduce on-the-job-injuries. The first study employed biomechanical modeling to determine a cutoff score for lifting tasks. The second study involved biomechanical analysis to determine the forces acting upon the knee joint and the relationship of the technique to knee injuries.

## **Validation of Force Required to Lift a Gurney**

The first study involves the development of a mathematical model to predict the force required to lift a patient-loaded gurney into an ambulance. It addresses the Paramedic job and the ability to lift a stretcher/gurney containing a patient in a safe and effective manner. The next slide illustrates one of the lifting tasks performed by the Paramedic. The gurney must be lifted off the ground to release the safety latch. The model was developed with the express purpose of determining the passing score for a physical performance test.

A detailed job analysis was completed to identify the critical physically demanding tasks. These tasks range from lifting and placing a patient into or out of the ambulance, lifting patients on backboards, and ascending/descending stairs in confined stairwells to external chest compression and others. However, two of the tasks with the highest criticality were: lifting of the stretcher from the collapsed to extended position and lifting it to a position to release the safety latch so that it can be rolled into the ambulance.

It was hypothesized that the minimum forces required to accomplish these tasks could be determined and validated against actual on the job performance of incumbent Paramedics. Thus the model objectives were derived. During the job analysis an ergonomic study of the workplace was conducted. Parameters such as heights, weights, body position, and equipment design were gathered. The critical lifting heights for the stretcher were 32 in. and 38 in. The ergonomic study also showed that the patients were predominantly male weighing between 191 and 216 lb. The stretcher specification were as follows: weight-63 lb., length-75 in., maximum height-32 in., and minimum height-8 in.

**HUMAN PERFORMANCE SYSTEMS, INC.**

A center of mass (COM) model was used to determine the force required to lift the head and foot ends of the stretcher. The COM is that point in the human body about which all parts are equally distributed. The assumptions used in the model included: (1) a patient weight of 200 in.; patient height of 71.37 in.; center of mass location of 56.18% from the soles of the feet which is the average for males (Cooper & Glassow, 1976); length of the stretcher bed, 75 in.; length of the handles of the stretcher of 0.75 in.; and the patient centered on the stretcher (the stretcher was equally balanced).

The equation was  $F_1 \cdot FA_1 = F_2 \cdot FA_2$  where:

$F_1$  =force at head end of stretcher

$F_2$  =force at foot end of stretcher

$FA_1$  =force arm length for head end of stretcher

$FA_2$  =force arm length for foot end of stretcher

The total equation is shown in the next slide. This model yielded a force of 150.7 lb. to lift the head end of the stretcher and 112.3 lb. to lift the foot end. Since Paramedics must share lifting each end of the stretcher, the force required to lift the head end was used in the remainder of the study.

An instrument was constructed to simulate the lifting of a stretcher with measurement taken at 32 in. and 38 in. If the lift was executed in a controlled manner a light would be activated. However, if the lift was executed in a forceful or jerky manner, a bell would be activated and nullify the lift. The machine was calibrated to determine the force required to lift a set of established weights (80–250 lb). The Paramedics lifted progressively heavier weights until they were unable to pass the 32 in. level. This type of lifting was similar to that encountered on the job.

A validation study was conducted to verify that these results were indicative of the minimum required by the Paramedics. The validation sample consisted of 15 female and 82 male Paramedics. To validate the model it was necessary to develop a criterion measure. Previous work by the research team (Gebhardt, Cooper, Jennings, Crump, & Sample, 1983; Gebhardt & Weldon, 1982) had shown that physical test scores for individuals in arduous jobs were related to supervisor ratings of (1) the worker's performance of physically demanding tasks and (2) the worker's overall job performance.

It was determined that supervisors did not view the physical aspects of the job and that the Paramedic's partners were the most knowledgeable about the physical aspects of job

performance. Therefore peer ratings were used as the criterion measure. The criterion measure consisted of a 6-point behaviorally anchored rating scale that defined varying levels of performance of specific job tasks. The job analysis and ergonomic study were used to select the tasks and develop the anchors. The criterion measure was pretested and revised. To ensure confidentiality the criterion measure was sent directly to the Paramedic's home and returned to the researchers in a prestamped envelope.

The results of the criterion-related validation study indicated that the criterion measure was reliable with an interrater reliability of .50-.66. The mean forces displayed by women and men were 162.5 lb. and 226.4 lb., respectively. The women's percent of the men's score was 71.8%. This percentage is on the upper end for women.

Correlational analysis, expectancy tables, and pass/fail ratios were used to determine the force that was associated with acceptable job performance for Paramedics. These analyses indicated that a force of 155.9 lb. was the minimum force associated with acceptable job performance. The force value of 150.7 lb. derived from the model closely approximated the force of value of 155.9 lb. obtained in the validation study. Therefore, it was concluded that the use of this type of center of mass model provided the minimum acceptable force for lifting a patient-loaded stretcher.

### **Analysis of the Descent Phase in Pole Climbing**

A major concern of women entering physically demanding work is that their job performance may be limited or they may become injured due to inadequate physical capacity. Although there are both males and females who do not possess the physical capabilities for specific physically demanding jobs, there is a lower percentage of women who can safely and effectively meet the standards in a job with high physical demands. Past research using basic strength tests has indicated that women possess 50% of the upper body strength of men and up to 85% of the lower body strength (Laubach, 1985). Research using work simulations has shown that the women's mean scores range from 50% to 60% of the men's mean scores (Gebhardt & Crump, 1989)

When women are confronted with a task that requires a high level of strength, they tend to compensate for lack of upper body strength by incorporating the use of the lower body musculature into the task performance. This is readily observed in lifting tasks. Women use the quadriceps, hamstrings, and gluteal muscles in the lifting process, while men tend to employ more upper body musculature (e.g., biceps brachii, pectoralis major and minor, etc.).

The key factor to insuring that women perform efficiently and safely in these jobs is to identify the nature of the requirements. Women are now found in the lineworker position in both the electric and telephone industries in the public and private sectors. One of the main tasks required in the job is the ability to climb the utility pole. This skill involves ascending a pole to a height of 20 to 100 feet while wearing climbers. The climbers (hooks) consist of long shanks with a stirrup at the lower end and a sharp gaff welded to the bottom of the shank. The worker straps the climber to the inside of the leg and uses the gaff to climb and descend the pole.

In the biomechanical analysis, male and female lineworkers were used as subjects to examine the effect of climbing technique on the forces at the knee joint. A review of medical records for outside plant jobs indicated that 15.4% and 19.3% of the accidents for males and females, respectively, involved the knee joint. Eighty-one percent of these injuries for women resulted from cut-outs (falls from the pole) or normal climbing procedures. Only 18.5% of the male knee injuries were related to pole climbing. These injury rates prompted investigation of the knee positions and the forces acting at the knee in both the ascent and descent phases of pole climbing.

The stability of the knee joint is dependent upon several musculoskeletal structures. The ligamentous structure consists of the medial and lateral collateral ligaments which are responsible for lateral stability and the anterior and posterior cruciate ligaments which are responsible for stability in the anteroposterior direction. Pain to the medial aspect of the knee was found to be the most prominent knee problem experienced by women. This complaint and further medical diagnosis indicated that there was medial collateral ligament involvement in the pain.

The initial comparison of the climbing technique of men and women focused on the path of the center of mass (COM). Inspection of the subjects' paths of the COM indicated that the patterns were quite similar. The greatest variation occurred in the length of the step taken during the ascent phase. However, this was attributed to individual technique. The circular pattern of the COM was due to the shift in weight from one foot to the other in the ascent phase.

During the ascent phase the maximum force (bone on bone) found at the knee was 281.8 lb. This occurred at the beginning of the push-off when one gaff was no longer in contact with the pole and most of the force was directed in a vertical direction.

During the descent phase the workers are instructed to extend the knee to a locked out position and allow gravity to pull the body downward to set the gaff in the pole. When the forces during the descent phase were calculated, the largest force (277.5 lb.) occurred at the point of impact with the pole, when the knee was in a fully extended position. Actually the knee passed



the extended position and hyperextended to 187° This position placed a great deal of stress on both the medial collateral and anterior cruciate ligaments due to the movement of the femur during extension. The force (bone on bone) at this point was almost double the weight of the climber. This force was in a posterior direction and resulted in a few additional degrees of hyperextension.

Prior to examining the full extent of the forces upon the knee joint, the male and female anatomical structure at the hip was reviewed. The angle of insertion of the head of the femur into the acetabulum differs for males and females. The 90° angle of inclination of the female femur is due to the wider, shorter, and deeper dimensions of the female pelvis (Steindler, 1955). This causes more strain on the medial collateral ligament in women than men. The male angle of 125° allows for a more vertical alignment of the shank and thus less tension in the medial collateral ligament.

The increased hyperextension discovered in the biomechanical analysis of the descent phase coupled with the increase angle at the hip yielded a higher resultant force at the knee in the medial direction for women than for men. This increased force was thought to be contributing to the increased ligamentous tears and general strain to the medial collateral and medial capsular ligaments of the knee.

The study resulted in recommendations that the descent technique be altered from a straight leg drop onto the gaff to one in which the knee is slightly flexed. This change in technique would reduce the impact force and change the direction of the resultant force to one that is less medially directed for women. This would decrease the injuries and worker compensation costs to a great degree. For example, the cost of a knee injury not only includes the surgery at a cost of \$1,500-2,000, but encompasses costs related to disability, claims processing, rehabilitation, and absenteeism. These costs typically total from \$5,000 to \$10,000 depending upon the severity of the injury (e.g., removal of cartilage, removal of meniscus).

## Summary

In summary, biomechanical modeling and analysis can be used effectively to determine valid job-related cutoff scores for physical performance tests. The key to modeling or measuring the actual forces or demands is dependent upon an understanding of the job tasks, the desired outcome of the tasks, and the mechanical and physiological parameters associated with task performance. Likewise, this understanding will facilitate the evaluation of task performance in relation to on-the-job injuries. Such analyses will help to identify the problems associated with task performance in relation to the mechanical and physiological makeup of the worker.

Results from such analyses can be used to recommend alterations to task performance techniques and in turn reduce injuries and worker compensation costs.

References available on request.

HUMAN PERFORMANCE SYSTEMS, INC.

192

-477-

# Ergonomic Principles and the Development of Physical Ability Standards

Oscar Spurlin Ph.D.  
ERGOMETRICS Inc., Seattle, Washington

T. L. Doolittle Ph.D.  
Department of Environmental Health  
University of Washington

## Introduction

This paper is intended to review research findings in the field of ergonomics and exercise physiology which are relevant to development of physical ability standards. Development of physical ability standards clearly requires a multi-disciplinary approach, yet most personnel specialists charged with this task have little exposure to the extensive literature from other fields which can be of assistance in development of fair and accurate tests. Three topics will be covered:

1. Defining Strength and Stamina,
2. Setting job-related standards based on a safe margin of reserve capacity,
3. Testing format.

The focus is on the principal dimensions of muscular strength and stamina. The relationship between an individual's maximum capacity and the safe performance of submaximal tasks is outlined. Finally we present arguments for the use of standardized or laboratory type physical ability measures in lieu of work sample or performance tests. Issues of safety, reliability, and accuracy are discussed.

## Strength and Stamina Defined

These two dimensions, exertion of force and generation of energy, are the primary ones for expressing an individual's capacity for meeting physically demanding job requirements. The first, the ability to exert force, is termed strength. The second, the ability to generate energy, is termed stamina.

## Strength

Strength means having sufficient muscle mass to exert enough force to move and handle the objects required by the job. Lifting, pushing or pulling is frequently encountered by all positions studied. The force required to manipulate the various objects was found to be from a few pounds to in excess of 100. It has been established by several investigators that the closer the weight of an object is to an individual's maximum capacity the more prone the person is to incurring a strain-type injury. Also, efforts that require more force increase the perceived difficulty experienced by the individual.

## Stamina

Stamina represents the body's ability to utilize oxygen to produce energy in sustaining prolonged repetitive activity. It is known technically as cardio-respiratory endurance, maximum aerobic power, or aerobic capacity. Metabolic demand can be expressed as kilocalories (food Calories), per hour, amount of oxygen consumed, or METs. METs is an expression of multiples of resting metabolic rate.

The greater the individual's stamina, the higher the energy cost of an activity that can be tolerated. Conversely, the higher the energy cost of an activity, the greater stamina that an individual must possess to successfully perform. As with strength, the higher the metabolic demand, particularly as it relates to the person's capacity, the greater will be the perceived difficulty.

## **NIOSH Guidelines for Strength and Stamina**

The National Institute for Occupational Safety and Health (NIOSH) has established guidelines for determining when control measures need to be implemented with respect to physical demands being placed upon the workers. Both force (strength), and metabolic (stamina), demands are considered. NIOSH has set two limits predicated upon expected strength and stamina capacities in the industrial worker population. The lower and higher levels are termed respectively, the action limit and the maximum permissible limit. According to NIOSH, demands that approach or exceed the action limit require administrative controls (e.g. pre-assignment or pre-employment screening), to insure that individuals will not be over stressed by the job. Demands that approach or exceed the maximum permissible limits require engineering controls (e.g. mechanical assist), or job redesign.

## **Indices Based on Organizational Experience**

The evaluation of job demands in keeping with the NIOSH guidelines requires formal ergonomic task analyses. Managers, however, first may perceive the need for such evaluations because of higher than desirable injury rates, turnover, absenteeism, or employee complaints. These are important signals that the job demands exceed the physical capacities of many in the workforce.

## **Establishing Physical Ability Requirements**

Because the scientific rationale underlying the development of physical ability tests is relatively new and evolving, the following sections provide more detail on the specific rationale used in analyzing the important dimensions of strength and stamina.

## **Determining Energy Costs of Job**

Several procedures may be utilized to directly assess the energy cost of job tasks. The first procedure entails monitoring heart rate of individuals actually performing the tasks and comparison to baseline data. This procedure is outlined as follows:

1. Submaximal exercise regimen on a step bench while connected to a heart rate monitor. This establishes each individual's capacity for work in terms of oxygen consumption as determined through heart rate. Figure 1 below illustrates this process.
2. Work sampling of key tasks while recording heart rate.
3. Information generated in step 1 is used to extrapolate the actual energy expenditure associated with each task for that individual. Figure 2 illustrates determining energy cost of a task based on two individuals.
4. Statistical analysis of all individuals, each performing the same task.

## **The Ergonomic Model for Predicting Energy Cost**

Given a sufficient data base of known energy costs for material handling jobs developed by the method described above, Garg, Chaffin and Herrin (1978), developed a methodology for the prediction of metabolic rates for material handling jobs. Much of the following description is from Garg et al (1978). The reader is referred to this reference for further discussion of the model.

"The Model is based on the assumption that a job can be divided into simple tasks (activity elements), and that the average metabolic energy expenditure rate can be predicted by knowing the energy expenditures of the simple tasks and the duration of the job. By dividing the job into tasks elements and assigning a metabolic cost to each task based on measurable force, distance, frequency, posture, technique, gender, body weight and time within each task, an energy requirement to perform just that task can be determined. The average metabolic energy requirement is simply equal to the sum of the energy demands of the task, and the maintenance of body posture, average over time." (Garg et al, 1978)

Garg et al (1978), further stated that: "The model validation showed a correlation coefficient of 0.95 between the measured and predicted metabolic rates. The coefficient of variation



(standard error/sample mean), was 10.2%". In other words, the methodology is highly effective for determining the metabolic costs for almost any industrial task. Before accepting this model, we compared the results it yielded with values of caloric costs for similar tasks reported in the literature, our own data base, and directly with those obtained from indirect estimates from heart rate response. In all cases the agreement between methods was excellent. Careful analysis did reveal that some of Garg et al's variables were more important than others in calculating metabolic cost. They can be classified into three categories in terms of their effects.

1. Greatest effect - body weight, frequency, and technique (i.e. type of lift -- squat, stoop, or arm),
2. Medium effect - distance, and measurable force or load,
3. Least effect - posture (i.e. standing or stooped), and gender .

time, which also is listed by Garg et al, obviously is important, but it becomes a common denominator, or baseline, upon which the other effects are compared. Body weight has a significant effect upon total energy cost, but when energy expenditure is normalized for size its relative importance diminishes.

A special comment regarding gender is appropriate because of affirmative action concerns. The literature on gender differences in energy costs for the same activity is mixed, and a complete discussion is beyond the scope of this report. The preponderance of studies report no difference because of gender. Garg et al (1978), in contrast, found differences for some of the movements as reflected in the corresponding equations. While the differences were of "statistical significance", they were slight and of questionable practical significance. In other studies, for example, differences in caloric expenditure attributable to gender, have been measured to be 4%. This is well within the 10.2% coefficient of variation expected with the method. This difference is mitigated further when costs and/or selection standards are normalized for body weight. While gender differences can be computed with the equations,

they appear to be of no practical significance: thus, no gender differences in screening standards resulted from this investigation.

### Determining Strength Requirements

The job demands can be translated into units of force and energy required to accomplish the tasks. In manual material handling tasks (a global term that covers most physically demanding jobs), the human provides the force and the energy. While the capacities to exert force and generate energy vary significantly among individuals, their efficiencies in doing so are reasonably constant. For example, lifting a 50 pound weight from the floor to a table requires the same force, and joint lever actions, regardless of the individual. Similarly, the energy costs for carrying 50 pounds up stairs at a vertical rise of 4 flights per minute would be expected to approximate 0.15 kcal per kilogram of body weight, regardless of the individual.

Individuals working near their strength capacity are three times more likely to suffer musculo-skeletal strain type injuries. Ayoub, Selan, and Jiang (1984), reported that the critical point approximates 75% of capacity for occasional lifting (a few times per hour). They further delineated normative percentage of maximum for repetitive day long lifting that individuals could lift safely. Thus, by knowing the weights that have to be lifted and the frequency that they have to be lifted, maximum strength requirements are determined so that individuals can safely perform the jobs. Figure 3 below provides an illustration of necessary reserve capacity based on task repetitions per minute.

### Summary of Research Procedure to Establish Physical Ability Requirements

The purpose of the job analysis is to assess the major physical capacities necessary to perform critical job tasks. From these assessments a job related screening for new hires is developed. Accordingly, we identify important tasks that are considered to be critical to job success which are also physiologically limiting. Panels of subject matter experts provide initial input with follow-up of a larger sample utilizing questionnaires

provide documentation of critical tasks. These are confirmed through work observations. The research conducted on the work loads associated with many jobs can be supplemented by the substantial data base available from other similar jobs.

The important research components contributing to analysis of physical abilities include:

1. Interviews with supervisors and job incumbent panels to define the critical tasks and perceived physical demands.
2. For large incumbent populations, a written job analysis survey of a representative sample is conducted.
3. Measurement of material handled by incumbents for dimensions and weight.
4. Observation of work being performed to determine number of task repetitions per minute, number of pieces handled per repetition, and distance over which material is moved.
5. Translate movements involved in critical tasks into a standard exercise test, such as curls, floor to knuckle lift, military press, etc. Determine maximum strength requirements for individuals to be able to perform material handling tasks without undue risk of injury. Establish sub-maximal lifting requirements on the test.
6. Videotaping of samples of key tasks while they are performed. Ten to twenty minutes of video is taken for each task. Tasks selected are those which were most important in determining the overall aerobic (stamina), demands of these jobs.
7. Analysis of movements recorded on videotape via the ergonomic model (explained above), to determine the metabolic requirements of the studied positions.

## **Establishing Safe Strength and Stamina Testing Formats**

### **Strength Test Format**

Ayoub et al (1984), developed a model for estimating shift long lifting capacity for varying frequencies (i.e. lifts required per minute), as a function maximum acceptable lift. This was an effective model, except that determination of maximum acceptable lift for any individual is very time consuming and subject to error. Doolittle and Kaiyala (1988), demonstrated that an individual's maximum strength capacity could effectively and safely be estimated from repetitive lifts of a submaximal weight. Doolittle (1989), presented a modification of the Ayoub et al model, that when combined with the Doolittle and Kaiyala equation, enables one to predict day long lifting capacity from a short term effort with a submaximal weight. Figure 4 below illustrates this equation.

Strength standards should not exceed the limits typically recommended in ergonomics literature. This research recommends 75 percent of an individual's maximum strength capacity for occasional efforts; 15-22 percent of capacity for continuous effort (see Figure 3).

While job requirements are conveniently expressed in terms of maximum capacity, actually testing for maximum subjects the applicant to a higher risk to injury than is warranted. In the ergonomics literature, maximum strength is often determined by trial and error, adding or subtracting weight until a maximum effort is determined. In addition to the risk, the trial and error approach required to determine maximum strength is time consuming for the applicant and the company.

The discussion above is predicated upon establishing maximum strength capacity in the major muscle groups used in task performance. We recommend use of standard dynamic exercise movements for this assessment. Candidates move an object through a full range of motion repeatedly. A standard exercise movement, such as arm curls, squat lifts, or military press, is used which most closely approximates the work. Sub-



maximal loads are used to estimate maximum capacity as described in the previous paragraphs.

Alternative formats which are commonly used in pre-employment screening are:

Isometric strength tests, Maximum exertion on fixed strain gauge.

Work Sample Test, such as dummy drags, stacking boxes, or stair climbs.

Isometric testing eliminates the trial and error approach of determining maximum strength, thus fatigue is not a factor. It also is less time consuming, however, in order to be reliable great care must be exercised to insure consistent joint angles are employed and muscle group isolation (through restraint mechanisms), is accomplished. Thus testing can be difficult, and the maximum exertion required does increase the risk of a strain type injury.

Non-physiologists often view work sample tests as inherently more valid than standardized exercise tests. A principal advantage of work sample tests is the assurance that the same muscle groups are employed in the test as on the job. We would argue that standardized tests can be related to key tasks just as accurately. Exercise tests have other striking advantages, including increased reliability, safety, and generalizability.

Development of a meaningful scoring criteria for work sample tasks is often the biggest drawback. Often, time to completion of the work sample is utilized, even though this time limit is somewhat arbitrary and is typically not a standard for actual job performance. The need for training to achieve maximum performance undoubtedly lowers the reliability of such tests.

Furthermore, this approach makes it difficult to accurately assess the individual's capacity relative to the task being performed. Again the real issue is ability to perform the task safely, not the ability to perform the task once.

### Aerobic Test Format

In order to insure that an individual possesses adequate aerobic power (stamina), a standardized

aerobic capacity test is administered. There are three accepted methods for administering this type of test (listed below). The correlation between these tests is high, and for the purposes of industrial screening, any one of the following may be used.

The aerobic test recommended, whichever form is employed, is a submaximal effort. Maximum oxygen consumption is estimated, rather than measured. This approach is far safer for the individual taking the test. Also, it is administratively feasible in terms of cost to the employer and time for the applicant. An actual determination of maximum requires monitoring by medical personnel and is time consuming and very expensive. The submaximal test when administered as recommended correlates very high ( $r = 0.90$ ) with actual determinations.

There are three alternatives, depending upon available equipment and administrative practicalities:

**Step Test:** Heart rate is monitored while the person steps up and down from a bench 0.28-m in height (11 inches) for six minutes. Stepping rate ranges from 18 to 24 ascents per minute. Heart rate, step rate, age, gender, and bench height are used to compute the individual's maximum aerobic power in mL/kg/min (American College of Sports Medicine, 1980; Astrand and Rodahl, 1986).

**Cycle Ergometry Test:** Heart rate is monitored while the person cycles on a stationary cycle at a known work load. Rotations per minute remains constant. The workload is increased after the first 3 minutes. Heart rate, age, gender, weight, and workload are used to compute the individual's maximum aerobic power in mL/kg/min (American College of Sports Medicine, 1980; Astrand and Rodahl, 1986).

**Treadmill Test:** Heart rate is monitored while the person walks/jogs on a motorized treadmill at a known speed and inclination. The workload is increased after the first 3 minutes by increasing the inclination of the treadmill. Heart rate, age, gender, weight,

and treadmill speed and inclination are used to compute the individual's maximum aerobic power in mL/kg/min (American College of Sports Medicine, 1980; Astrand and Rodahl, 1986).

### Guidelines

Some of our experience in designing physical ability tests might be summarized into these guidelines:

1. The first guideline is that the specific muscle groups employed on the job should be the ones tested.
2. The second guideline is that tests should be objective and reliable indices. Candidates should be able to achieve their maximum score without significant training or coaching.
3. The third guideline is that strength requirements should not exceed the limits typically recommended in ergonomics literature. This research recommends 65-75 percent of an individual's maximum strength capacity for occasional efforts and 15-20 percent for repetitious work.
4. The fourth guideline is that the tests should be made as safe as possible, given the untrained nature and unknown medical history of job applicants. It is possible to accurately estimate maximum dynamic strength from repetitions requiring a sub-maximal (and safer), effort.
5. The fifth guideline relates metabolic demand to maximum aerobic power, where it is well recognized that individuals cannot be expected to perform short term efforts at greater than 75-85 percent of maximum or day long efforts at greater than 33-40 percent of capacity, (Astrand and Rodahl, 1986).

### References

American College of Sports Medicine, Guidelines for Graded Exercise Testing and Exercise Prescription, Lea and Febiger, 1980.

Astrand, P. and Rodahl, K., 1986, Textbook of Work Physiology, 3rd edn, (New York: McGraw-Hill).

Ayoub, M., Selan, J., and Jiang, B., 1984, Mini-Guide for Lifting, (Lubbock: Texas Tech University).

Doolittle, T., 1989, Aerobic testing for manual handling Jobs. In: Advances in Industrial Ergonomics and Safety I, Mital, A., (Ed). (London: Taylor & Francis).

Doolittle, T., and Kaiyala, K., 1988, Prediction of maximum dynamic strength from multiple repetitions with a submaximal load. In: Trends in Ergonomics/Human Factors V, Aghazadeh, F. (Ed), (North-Holland, Amsterdam: Elsevier), 767-774.

Garg, A., Chaffin, D. B., and Herrin G. B., 1978, Prediction of metabolic rates for manual materials handling jobs, In: American Industrial Hygiene Association Journal, vol. 39, pp 661-74.

NIOSH, 1981, Work Practices Guide for Manual Lifting, (Cincinnati: National Institute for Occupational Safety and Health, publication number PB82 178948).

Figure 1

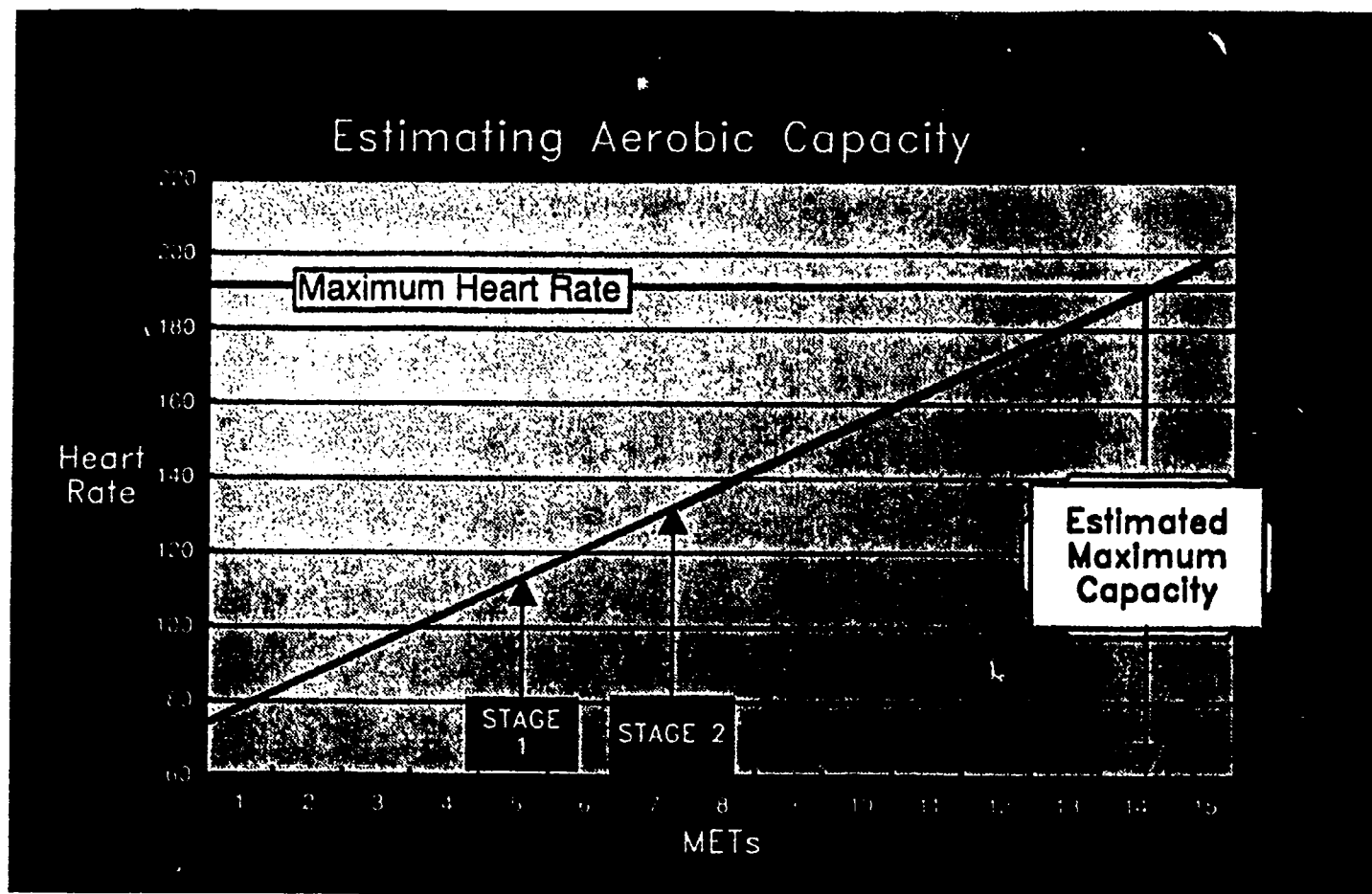


Figure 2

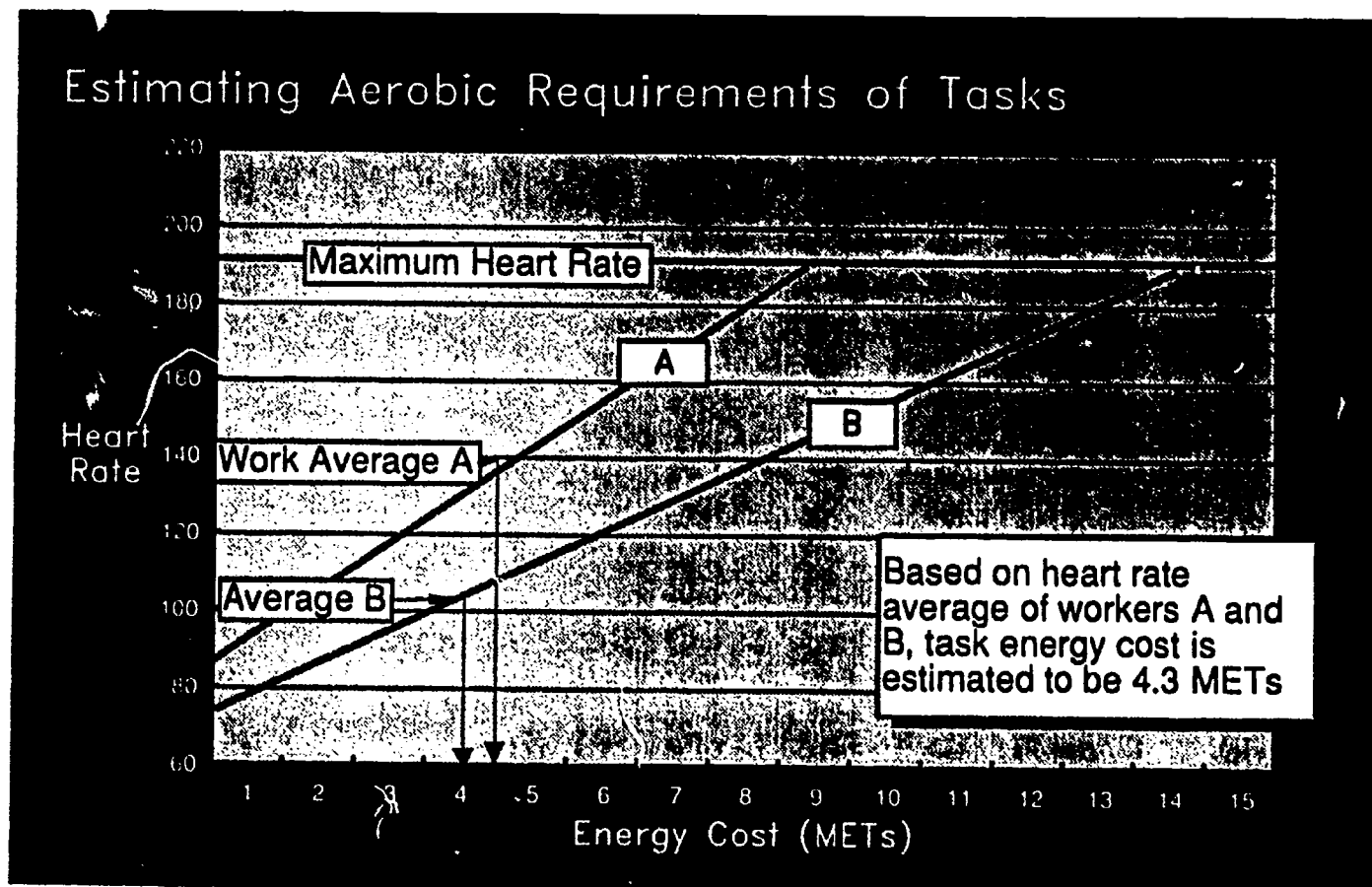


Figure 3

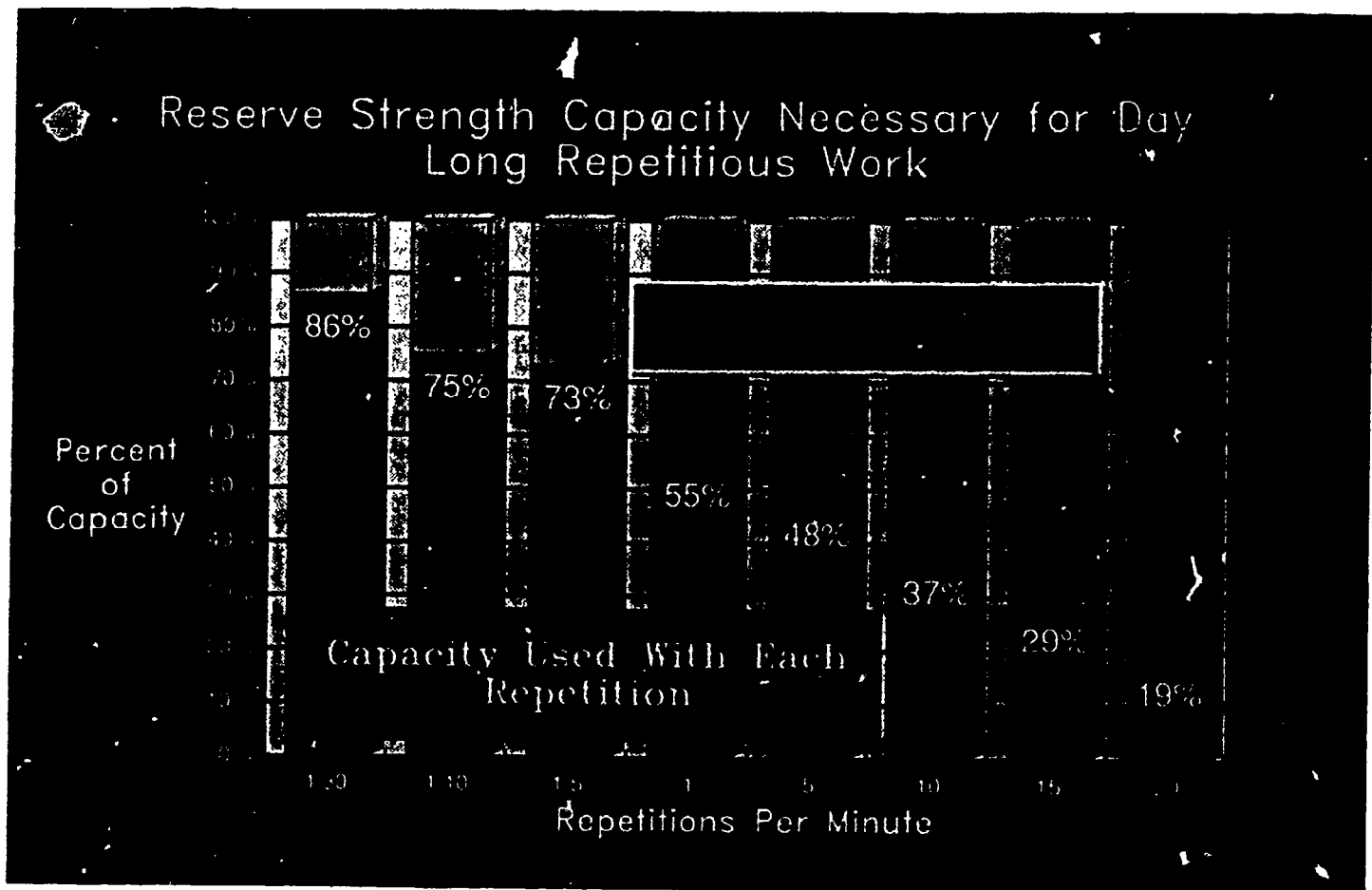
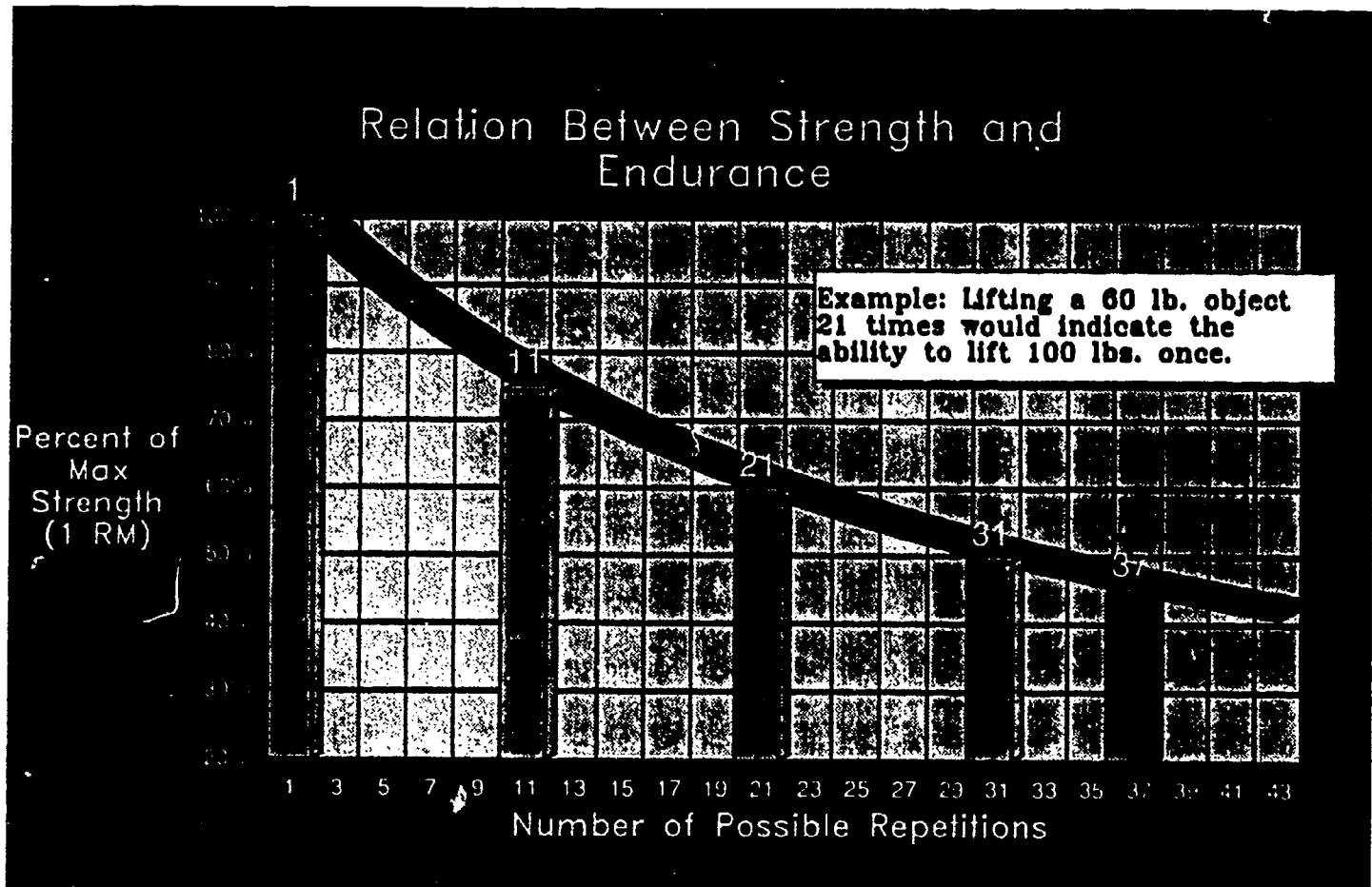


Figure 4



**JEANNERET & ASSOCIATES, INC.**

**MANAGEMENT CONSULTANTS**

P.O. Box 200039  
Austin, Texas 78720-0039  
(512) 250-5156

3223 Smith Street, Suite 212  
Houston, Texas 77006-6685  
(713) 529-3015

P.O. Box 12007  
St. Louis, Missouri 63112-0107  
(314) 862-3805

**JOB QUALIFICATIONS LINKAGE SYSTEMS**

**Symposium Prepared for the  
1990 Annual Conference on Personnel Assessment  
of the IPMA Assessment Council**

**Assessing Workforce 2000**

**June 28, 1990**

**S. Morton McPhail, Ph.D.  
John R. Moore, Ph.D.**

**Jeanneret & Associates, Inc.**



## Introduction

A common problem encountered by employers who must process large numbers of applicants for a variety of positions is that of effectively and efficiently screening applications prior to more formal and complete assessment of applicants' capabilities. An approach which is often taken in conducting such an initial screening is to establish a set of minimum qualifications for each position. Some employers also establish a set of desirable qualifications which go beyond the minimal requirements for entry into a position and are used to flag particularly well-qualified applicants. This process of relying upon pre-defined minimum qualifications that are largely objective in their application and the subsequent definition of somewhat more subjective desirable qualifications greatly enhances the effectiveness of the initial screening and reduces the cost and enhances the efficiency of evaluating the remaining applicants.

The difficulty that arises with this process is that there is a clear need to support the screening criteria, both internally to the organization and externally as well. Human resource professionals often find themselves serving a dual role as both the enforcers of rules and regulations and the enablers of management to ensure quality selection. As enforcers, they have the obligation to ensure and document the job-relatedness of requirements in compliance with relevant equal employment opportunity enforcement guidelines and to oversee the fair application of those requirements to all candidates. As enablers, they have the mandate to assist management in an efficient manner in identifying the best available candidates and to ensure that the qualifications established will select candidates capable of performing the required work activities.

From an internal viewpoint, it is frequently necessary to document and define the entry level qualifications established for a position to the managers and decision makers for whom referrals of applicants are to be made. It is not uncommon for managers to seek to establish minimum qualifications and even desirable qualifications which meet some particular purpose for their work unit or organization. For example, it is not unknown for managers to establish criteria which excessively narrow the field of applicants who can fulfill them. Alternatively, a confusion may exist between requirements that are functionally minimal to a position, and justify *a priori* de-selection of applicants, and those requirements which are useful but not absolutely essential to a position. Thus, managers may in a good-faith desire to improve the effectiveness of their work



force establish minimum qualifications which appear to be excessive for initial screening of applicants.

On the other hand, criteria for selection must be sufficiently restrictive to ensure that those applicants who are referred for further processing do, in fact, have the minimum capabilities to accomplish the work. These needs place a strong responsibility on human resource professionals to find that narrow line between requirements which are sufficiently stringent to give some assurance of the capabilities of selectees while simultaneously not overly restrictive of the applicant pool, both from the standpoint of recruiting and for meeting other organizational goals. Managers and human resource professionals share responsibility for ensuring that the requirements established are, in fact, relevant to the job requirements in such a way that the quality of the selections made will be acceptable and defensible.

The screening criteria must also be defensible from an external viewpoint. Especially in the public sector, it is frequently the case that an organization's public relations suffer when screening criteria are established that appear to be unreasonable or inadequate to select appropriate personnel to fill positions. Beyond the simple appearance of propriety, however, is the legal constraint placed upon all employers by the requirements of the federal and state equal employment opportunity enforcement agencies. Much case law and administrative precedence establish that any selection methods, including the creation of minimum qualifications, must be demonstrably job-related. Any requirements which are used as the basis of employment decisions are subject to the same validation requirements as any written test or other employment process. Thus, the minimal qualifications which are used to screen the initial applications, as well as the desirable qualifications which may be determinative in selecting among applicants who meet the minimum qualifications, are subject to the same requirements of documented job-relatedness as any other selection procedure.

A difficulty arising with respect to these requirements, however, is the expense in terms of efficiency and time commitment which arises in trying to conduct extensive validation analyses. Indeed, some of the requirements which are commonly established as initial screening criteria (e.g., educational requirements) are generally not amenable to traditional validation techniques and frequently encounter the twin difficulties of inadequate numbers of people and inadequate

time and resources. Thus, on the one hand are the reasonable requirements to demonstrate job-relatedness of screening criteria as well as the legal obligations to do so in order to ensure appropriate and effective selections are made. On the other hand, there exist constraints on the analyses engendered by organizational realities.

The purpose of this symposium is to present an alternative approach to developing the documentation required to demonstrate the job-relatedness of qualifications used for initial applicant screening. The technology to be demonstrated will also offer insights and assistance in developing more sophisticated selection methodologies and can serve as the first step in searching for appropriate alternatives for pre-employment testing. This methodology can provide an enormous amount of information efficiently and effectively while organizing that information and putting it in an effective and usable format.

### Job Analysis

All efforts to document and support the use of any selection requirements necessarily begin with an analysis of the work activities. Job analysis information comes in a variety of forms and levels of detail. Two broad categories of job analysis can be identified. The first of these is job-oriented information; most of us would recognize job-oriented information as traditional task analysis. A second broad category can be characterized as worker-oriented job analysis information. Worker-oriented job analysis information describes jobs in terms of the demands placed upon the worker by the job activities. That is, the job is described in terms of the human behaviors or attributes required to perform the task.

These two approaches to job analysis have, for many years, found complimentary use as part of research to demonstrate the validity, or job-relatedness, of selection requirements. One of the salient features of the methodology to be described today is a linking of job- and worker-oriented information. In many ways, the power of this technology lies in the establishment and documentation of exactly that linkage. In some ways this linking of information is not different from the linking of knowledge, skills, and abilities with job tasks which is frequently a part of test development and validation efforts. However, by using detailed worker-oriented job analysis

information, the process is more precise and more fully documented than otherwise would be possible.

Starting with job analysis information, the process of demonstrating the validity or job-relatedness of job selection requirements generally follows two main strategies, both of which are recognized professionally and legally. These are, of course, the content and criterion or empirical validation strategies. Differentiating these two should not be construed to imply that they are not related to one another. Rather, differentiating them is helpful primarily in describing the kinds of information which may be brought to bear to support the job-relatedness of selection criteria or qualifications. Technically, content validation is the process of specifying the overlap between a selection criterion and the content domain of the job. The criterion-related validation strategy relies upon a statistical demonstration of the co-variation of the selection qualification and some measure of job performance. The methodology of the Position Qualification System (PQS) owes its genesis to both content and criterion strategies and to a linking of the strategies via conceptual and statistical methods.

The statistical methods used allow for the linking of job information to people information. By examining the tasks performed on a job, the human demands requirements placed upon the person to perform those tasks, and a large base of information regarding successful measures of those human requirements, estimates of the appropriate measures to assess the requirements can be obtained. Previous research allows for the linkage and documentation of the job activities with appropriate measures of human attributes.

It is not intended that all of these analyses can happen automatically based on standardized decision rules. Job experts must make judgments in applying this methodology (as in all job analytic approaches) about the human requirements necessary to perform the job activities. However, the conceptual procedures and automated software combine to provide for a logical and rational approach to documenting the job-relatedness of selection criteria. This methodology will not do all the work, but it can make the establishment of selection criteria more precise, more efficient, better organized, and more carefully documented.

As noted above, the starting point for developing qualification systems is job analysis in order to be able to demonstrate matches between tasks performed, knowledge, skills, abilities, and other characteristics (KSAOs) required, and applicant assessment procedures. Three types of job analysis data are necessary to establish these relationships: task information, worker behaviors or attributes, and human attributes required to perform the job activities. Unfortunately, most existing job analysis procedures result in only one type of data, and it has often been recommended that multiple types of job analyses be conducted simultaneously. Traditionally, however, multi-method job analysis has not lent itself to efficient storage, organization, or retrieval.

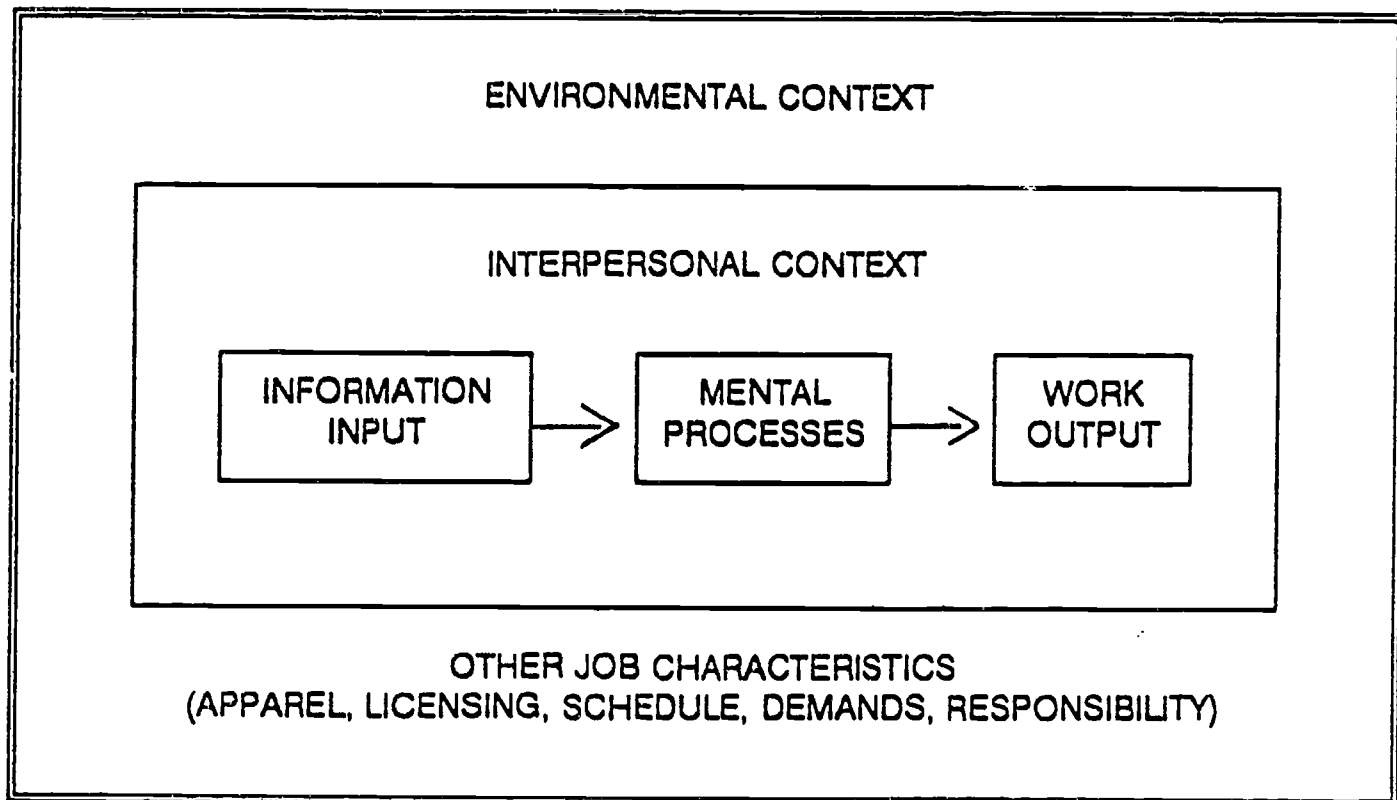
### Position Analysis Questionnaire

Using the technology of the personal computer, a method for efficiently combining task-, attribute-, and worker-oriented data is now available. The starting point is the Position Analysis Questionnaire (PAQ). The PAQ was developed beginning in the late 1960s at Purdue University as part of a long-term research program under the direction of Dr. E.J. McCormick. It was originally published in 1969. Since then, it has been used by over 1,000 organizations and continues to be used in and the subject of scholarly research. It has been cited in litigation as a model of job analysis. In 1979 it was cited as one of the most significant milestone in job analysis methodology during the last 60 years (Dunnette and Borman, 1979, *Annual Review of Psychology*). To date, almost 300,000 PAQs have been completed.

The PAQ is a structured, quantitative worker-oriented job analysis tool consisting of 187 job elements (items) that are rated numerically. These 187 elements measure six critical areas of jobs:

[INSERT FIGURE 1 ABOUT HERE]

FIGURE 1  
CONCEPTUAL ORGANIZATION OF THE PAQ



- **Information Input.** To perform any job, a person first needs to obtain information about what is to be done. This information can be gained from written sources, oral instructions, observation of materials, equipment, people, or events in the environment. The PAQ measures the extent to which a job requires the worker to obtain information from a variety of sources.
- **Mental Processes.** Once information is obtained, the worker "uses" or "processes" it to determine what actions to take. Mental processes include decision making, planning, analyzing, and compiling, as well as others, within the training, education, and experience requirements of the job. The PAQ measures how information is processed and the importance of processing to job accomplishment.
- **Work Output.** Once information is obtained and processed, jobs require the worker to perform some action or set of activities. This action may involve using tools or equipment, operating vehicles, handling materials, physical exertion, writing, or otherwise actively engaging in some work activity. The PAQ measures the activities performed as required by the job.
- **Relationships with Other People.** In the course of working, most individuals have some contact with other people relevant to job activities. It is important to examine with whom a worker interacts, what form the interaction takes, and what type of supervision or direction an individual provides to/receives from other workers. The PAQ allows for a quantitative description of how a worker communicates with others (e.g., advising, selling, providing information, etc.), with whom communication occurs (i.e., other positions or jobs), and the scope of supervisory responsibility.
- **Job Context.** Clearly, the environment in which an individual works has a bearing on his/her job activities. Physical working conditions include temperature, noise level, illumination, work space, hazards, and indoor versus outdoor location. Mental working conditions include stress, pressure, frustration, and interpersonal conflict. The PAQ measures the extent to which various working conditions have an impact on the job.



- **Other Job Characteristics.** Other factors related to work activities that do not fall into the previous categories include licensing requirements, specific job demands, responsibility for material assets and the safety of others, the amount of job structure, and the criticality of a job in terms of its impact on the organization's operations, assets, or reputation.

In addition to this worker-oriented information, previous research with the PAQ established links between the PAQ job element and 76 common attributes (e.g., mechanical ability, finger dexterity, dynamic strength, influencing people, and empathy). The initial research had experts in the field of job analysis rate the relevance of the 76 attributes to each job element in the PAQ. High levels of agreement, measured by the inter-rater reliability coefficient, were found between the experts' ratings. Extensive subsequent empirical research supports the links between the PAQ job elements and the 76 human attributes.

In summary, the PAQ provides comprehensive behavioral and contextual information about a job. Specifically, when a PAQ has been completed, it is scored and yields the following types of information about a job:

- The job elements that are especially relevant or important for the job.
- Work interests that are especially relevant to the job requirements.
- Aptitudes and abilities that are especially relevant to job performance.
- Education and experience requirements of the job.
- Predicted usefulness for the analyzed job of the General Ability Test Battery (GATB) subtests.
- Job Dimensions (groups of items) that are especially relevant to describing the job.
- Prediction of FLSA exempt status for the job.
- Job evaluation points.

### **Position Qualifications System**

Using this information, many general requirements can be identified directly from a "scored" PAQ. However, in order to demonstrate and fully document content validity, identify specific required job knowledges, and develop minimum and desired levels of qualifications, it is

necessary to match the PAQ information to tasks. Since the PAQ is automated, a software program (the PQS) has been developed that allows task information to be added to the job analysis data base. Further, the program allows the user to match tasks to relevant PAQ job elements, job dimensions, work interests, aptitudes, and abilities in order to identify which of these job characteristics are relevant to specific tasks and to quantify the degree of relevancy. This organization and analysis of the data facilitates the identification of essential KSAOs and their minimum and desirable levels. In addition, educational and experience requirements are provided, as are the most-likely-to-be-useful GATB sub-tests. Reports can be produced that provide documentation of these relationships and linkages. The PQS system is diagrammed in Figure 2. The PQS process is described in more detail below.

Step 1 of the PQS process entails direct input of a job brief and tasks into the existing job analysis data base. There also is the option to include "time-spent" and "importance" ratings for each task; the system will calculate "criticality" ratings for each task. The tasks can be obtained from job descriptions, task lists, or job analysis interview notes.

Step 2 of the process allows the user to rate quantitatively the relevancy of each task with (1) the most highly rated PAQ job elements; (2) the most highly rated human attributes; and (3) the most important rated job dimensions. In addition, educational and experience requirements are provided directly from the PAQ analysis. These latter requirements can be further tailored to the specifics of the job being analyzed.

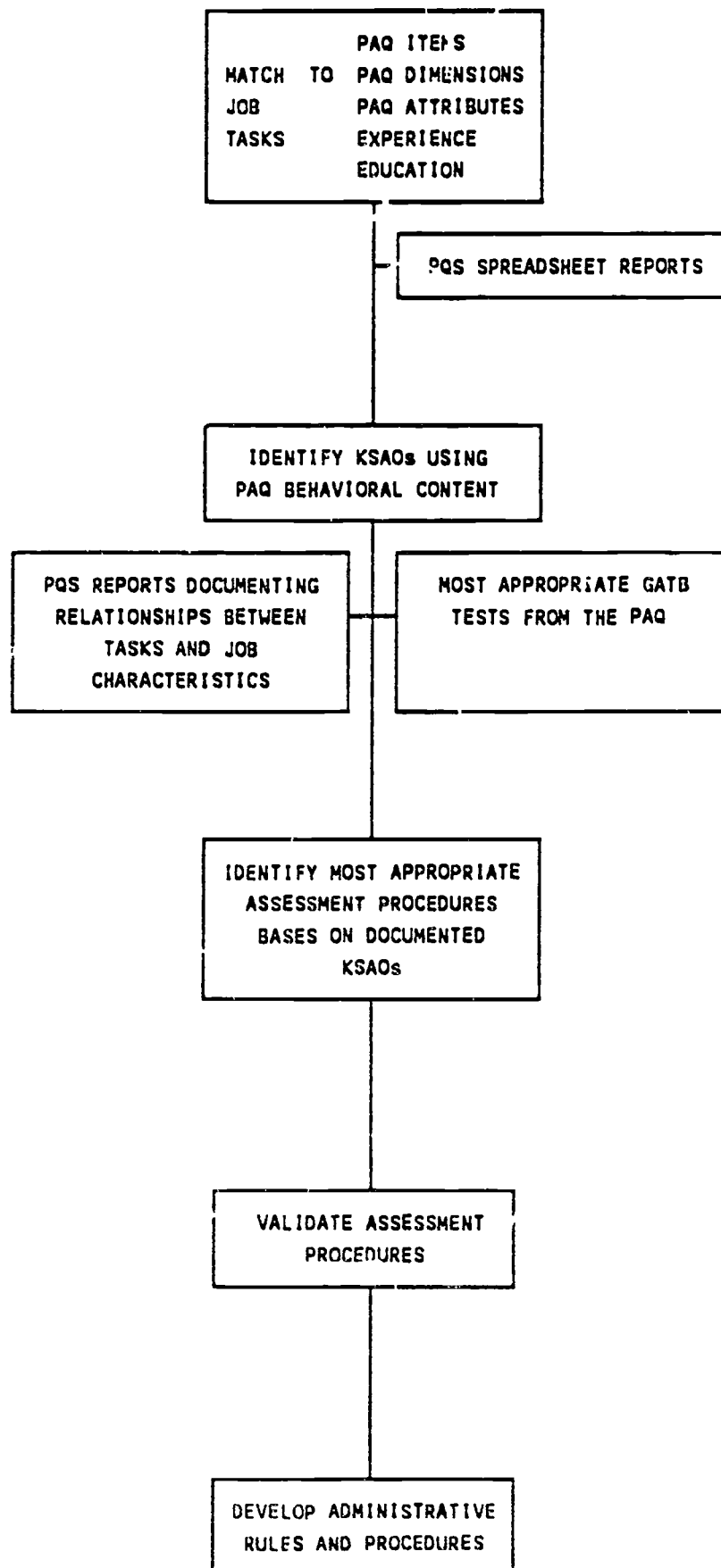
At the completion of this step, the system will generate spreadsheet reports that display the job elements, human attributes, and job dimensions relevant to each task. The education and experience requirements are provided as well.

Step 3 involves analysis of these reports to identify relevant KSAOs. For example, the reports can be examined to determine the job characteristics (i.e., job elements, human attributes, and job dimensions) that are matched to several tasks or the most critical tasks. Some of these job characteristics will represent specific KSAOs; some KSAOs may have to be derived or inferred from the pattern of job characteristics.

FIGURE 2

DEVELOPING PERSONNEL QUALIFICATIONS SYSTEMS

USING PQS



Step 4 of the process require that the relevant KSAOs be input into the data base and quantitatively rated for their relevancy to each task. A report displaying the relationship between tasks and KSACs can be generated at this point. This report provides complete documentation of the job-relatedness of the KSAOs and the foundation for the content validation evidence for the applicant assessment procedures.

The identification and development of procedures to assess applicants for the relevant KSAOs is relatively straightforward for human resource professionals with training or experience in applicant selection procedures. In particular, the predicted usefulness of individual GATB tests will be listed directly from the PAQ analysis. If GATB tests are unavailable, other tests that assess the same attributes can be substituted, subject to validation. Other types of assessment procedures that could be supported by a PQS analysis include structured interviews, knowledge tests, work simulations, physical ability testing, and personality assessment. Once likely applicant assessment procedures have been identified, they should be formally validated to ensure compliance with federal, state, and local regulations, as well as to ensure maximum usefulness. Again, the PQS analysis will supply the majority, if not all, of the necessary evidence to support the content validity of the assessment procedures. Additional steps, of course, will be required to establish criterion-related validity. The last step in the process would be to develop administrative rules and procedures.

Once this process has been completed for a job, the data are readily stored and easily accessible for future use. Subsequent hirings for a job would entail merely checking and updating the existing PQS information.

#### Assessment Implications for Workforce 2000

The nature of the workforce in the year 2000 is frequently predicted to be smaller than it is today, lacking in many essential skills, and relatively more diverse. In addition, a trend in business is to increase productivity through staff reductions, cost containment, and introduction of technology. This scenario suggests that there may be fewer human resource professionals available for assessing and selecting applicants. Automation likely will become

an essential tool for these professionals to develop and administer assessment programs cost effectively.

Perhaps more importantly, identification and selection of qualified applicants will become more difficult and critical if there are fewer applicants possessing required qualifications. Accordingly, the capability to identify quickly and accurately relevant KSAOs and corresponding assessment/selection procedures will enhance an organization's ability to hire the types of skilled personnel it needs.

Finally, automation of job analysis data can allow for the monitoring of the availability of different KSAOs in the workforce. For example, it would be possible to identify KSAOs that are difficult to find, or alternatively, KSAOs that are in abundance. This information is valuable in considering recruitment strategies, the design of jobs, the mix of jobs in the organization, and even the organization of work responsibilities. In essence, an automated job analysis system is a management information tool, providing the human resource professional with timely data to make informed decisions about the present and the future.

THE HIERARCHICAL JOB ANALYSIS:  
A STRUCTURED APPROACH TO THE JOB ANALYSIS INTERVIEW

David E. Smith  
Anheuser-Busch Companies, Inc.  
Human Resources Development  
and Selection

Fran Laue  
UES Incorporated  
Human Factors Division

Robert M. McIntyre  
Old Dominion University  
Department of Psychology



## The Hierarchical Job Analysis: A Structured Approach to the Job Analysis Interview

The fact that a thorough job analysis is essential to good personnel management goes without questioning. With accurate and complete information obtained from the analysis, an organization can successfully develop recruitment, selection, and training programs, establish performance appraisal systems and make meaningful comparisons between jobs for the determination of pay equity. The job analysis has become increasingly important with today's challenges for improved productivity in American industries. It has also become the focal point in the legal context of human resource selection (Gatewood and Field, 1987).

There are various methods for conducting job analysis. Direct observation, interview, job inventory, critical incidents and the Position Analysis Questionnaire are but a few of the techniques that have been employed over the years (Cascio, 1982). Often several approaches are combined in job analysis studies. One of the most commonly used technique is the job analysis interview (Rendero, 1981). A face-to-face interaction with job incumbents or other job experts provides the opportunity for a level of detail that is often unavailable through other methods.

Despite the potential benefits of the interview, the success of the technique depends heavily on the skills of the job analyst. The most important requirement of a job analysis is to insure that all of the essential aspects of the target job are identified. In addition, if job analysis results are to be useful, they must be

organized in a meaningful and systematic way. Unless a well planned approach to the interview is followed the job analysis can result in a fragmented and incomplete list of job information.

The job analysis interview can provide thorough and meaningfully organized results by insuring that the approach closely attends to the orientation of the job analysis and the level of detail used to describe the job. Jobs can be described using a task orientation (what gets done on the job) or a worker orientation (how the job gets done), or in terms of knowledges, skills or traits required to perform the job (Arvey and Faley, 1988). Without a unified orientation the results of the job analysis become less clear and may provide comparisons of job aspects that are misleading.

A job can also be described in various levels of detail, from the most global description of job responsibilities to specific actions or behaviors required on the job. To be useful, job activities should be organized on the basis of its level of detail. The risk of not doing this is to possibly place improper emphasis on certain aspects of the job.

### The Hierarchical Job Analysis

The Hierarchical Job Analysis (HJA) is a technique that provides direction for the job analyst in conducting the job analysis interview. It guides the analyst in identifying the essential aspects of the job and provides results that are organized in a meaningful and systematic way. The HJA provides specified levels of detail within statements describing the job

and maintains a unified orientation within these statements. It accomplishes this by using simple guidelines for conducting the job analysis interview.

The HJA could be thought of as an attempt to describe the job in terms of a "job-tree". Using a task orientation, the technique requires job experts to break down the target job into a number of sub-jobs that could be performed independently by two or more individuals. Each sub-job is further broken down into a number of independent tasks making up that sub-job. Again, each task is further broken down into a number of sub-tasks.

This successive process of partitioning the job into more and more specific units is continued until it is felt that further breakdowns would produce work behaviors that are not meaningful to the analysis. The outcome of this approach is represented in a dendrogram or job-tree. (See Figure 1 below.) The job analysis results resemble that of an inverted tree, with a small number of sub-jobs appearing near the top of the diagram and a larger number of sub-tasks appearing at or near the bottom.

#### A Case Study: The Operating Room Nurse

To evaluate the effectiveness of the HJA, the technique was used to analyze the position of Operating Room Nurse (ORN) at a northern Colorado hospital. At the time of the analysis approximately 30 Registered Nurses were employed by the hospital to function as ORNs. The surgical department had identified a need to revise their performance appraisal program for Operating

Room Nurses and Technicians. This served as a good opportunity to test the utility of the HJA in an applied setting. The job

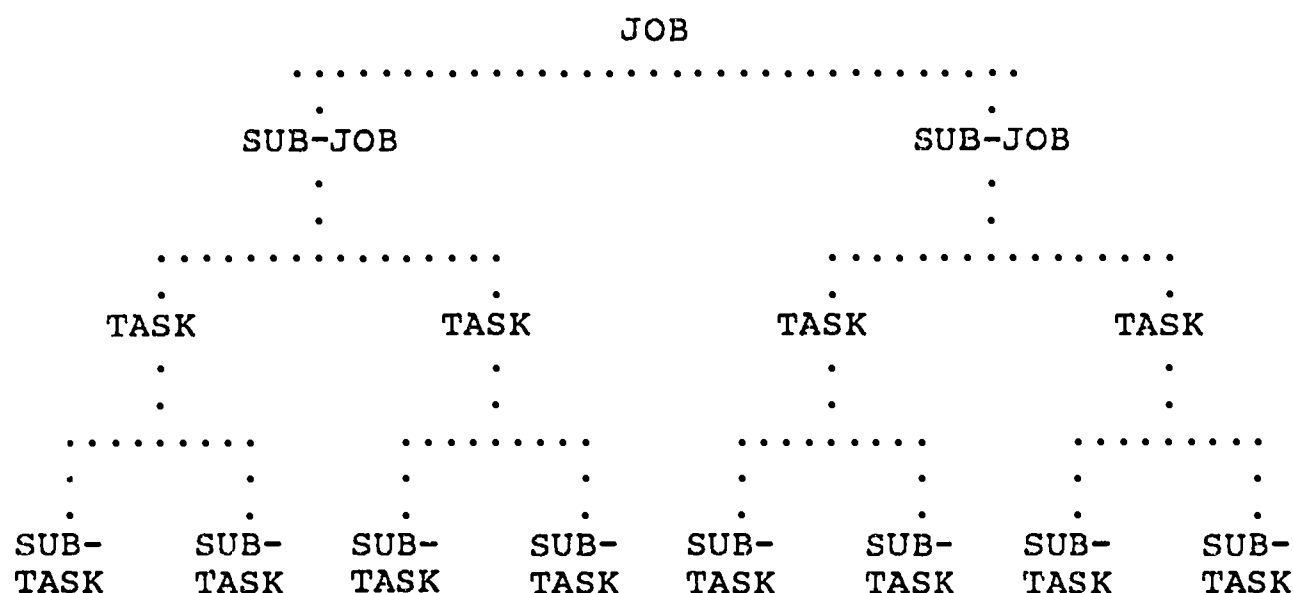


Figure 1. A sample job-tree diagram

analysis study followed three stages: (1) the training of a small group of job experts in the use of the HJA technique; (2) the development of a list of job activities by the job experts; and (3) the development of a job inventory that could be used to gain input from the entire nursing staff.

#### STEP 1: HJA Training

The job experts involved in the project were six nurses chosen from a list of ORNs who had been employed in their present positions for a minimum of one year and who were sufficiently familiar with all aspects of the job. The authors met with the

six job experts in a series of five ninety-minute meetings. At the initial group meeting a brief training session was conducted, providing an introduction to the purpose of the job analysis and describing the Hierarchical Job Analysis approach.

The training included a description of the goal of the process (i.e., complete coverage of the job, a unified orientation in the job statements, and specified levels of detail). In addition, the job experts were shown a sample job-tree of a familiar job to illustrate the type of product that was expected from the project. The job represented by the sample job-tree was that of a waiter or waitress in a coffee shop. The sample job-tree used in the training is presented in Figure 2. For demonstrative purposes, the job-tree included only a sample of the job activities of the waiter/waitress position.

#### STEP 2: Developing a List of Job Duties

Following the brief training period, the job experts began dividing their job into conceptually distinct sub-jobs. To aid them in this process, they were told to imagine that two or more equally trained nurses were provided to share in the duties of their job. They were then asked to break down the major functions of their job so that independent functions could be delegated to several nurses. Each job expert worked individually for a few minutes generating a list of sub-jobs. A group discussion then followed to compare and evaluate each job expert's list and to establish one final list of sub-jobs. The sub-jobs defined by the





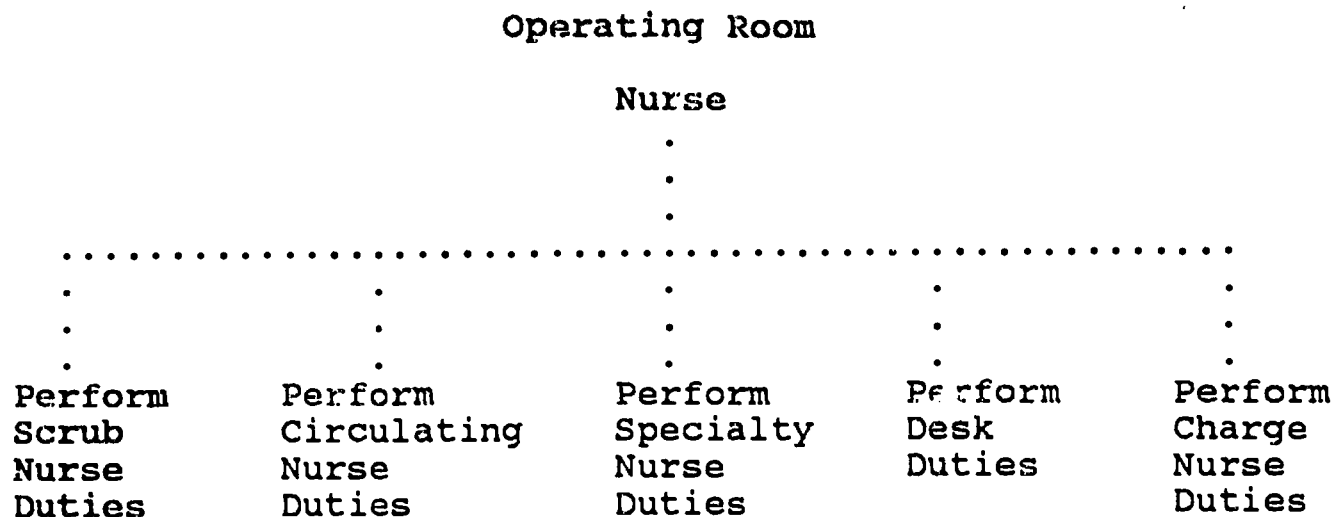


Figure 3. Five sub-jobs of the Operating Room Nurse

In the next step of the job analysis, the job experts were asked to divide the five previously identified sub-jobs into tasks that make up each sub-job. Focusing on one sub-job at a time, the job experts were asked to consider ways in which the sub-job could be divided into two or more smaller, independent tasks. Again, the nurses worked individually, generating a list of tasks. The entire group then discussed each person's list and came to a consensus on a final list for that sub-job. For example, in breaking down the sub-job of the scrub nurse, the job experts identified eight independent tasks that they perform, ranging from preparing the operating room for surgery to teaching surgical procedures to other personnel. The final list of tasks for the scrub nurse sub-job is shown in Figure 4. A total of 31 tasks were identified for all eight of the sub-jobs.

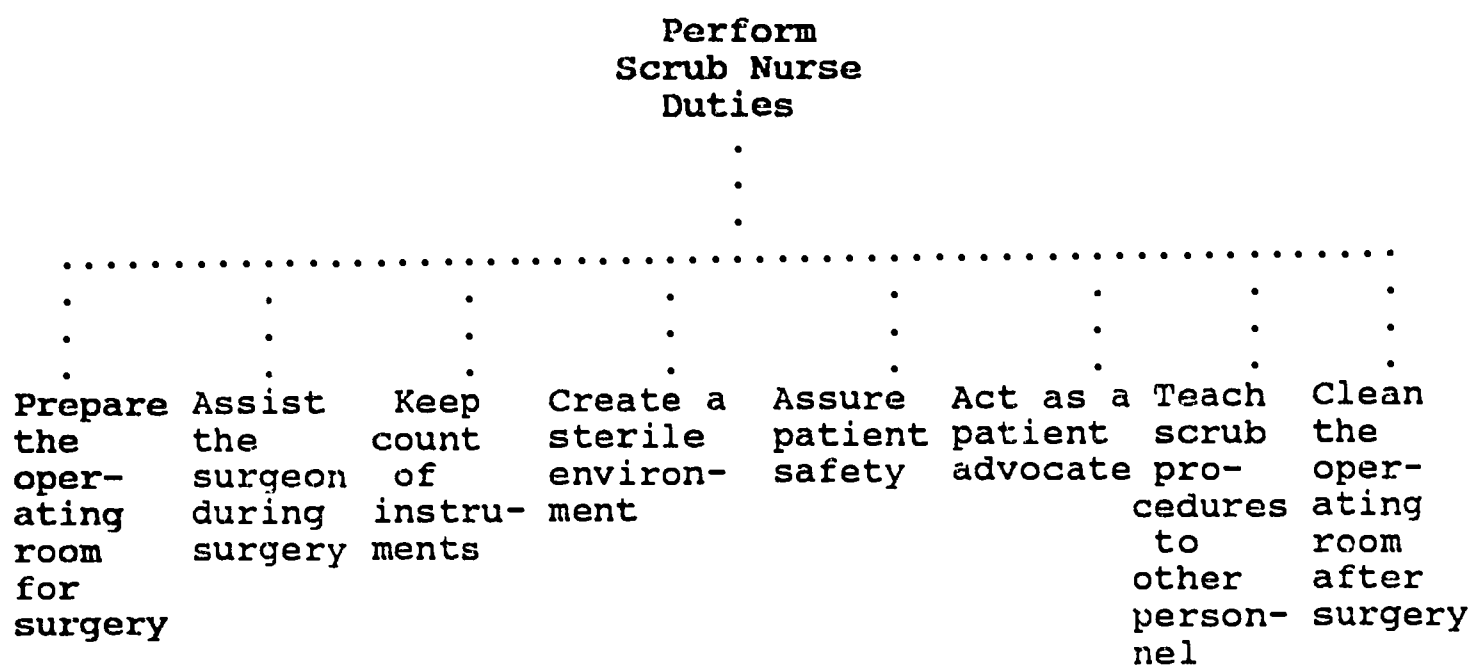


Figure 4. A breakdown of the Scrub Nurse sub-job

The next step in the job analysis required the job experts to break down each task into sub-tasks. Similar to the previous step, the nurses focused on one task at a time until breakdowns on all tasks had been completed. The 31 tasks of the ORN were ultimately broken down into a total of 161 sub-tasks. An example of how these tasks were broken down is provided in Figure 5. The task, preparing the operating room for surgery, was described by four sub-tasks.

At this point it was felt that the statements describing the sub-tasks were sufficiently detailed for the purpose of the job analysis and the process was discontinued. Throughout the entire process the job experts were reminded of the objectives of the job analysis -- to list and describe all of the essential duties of

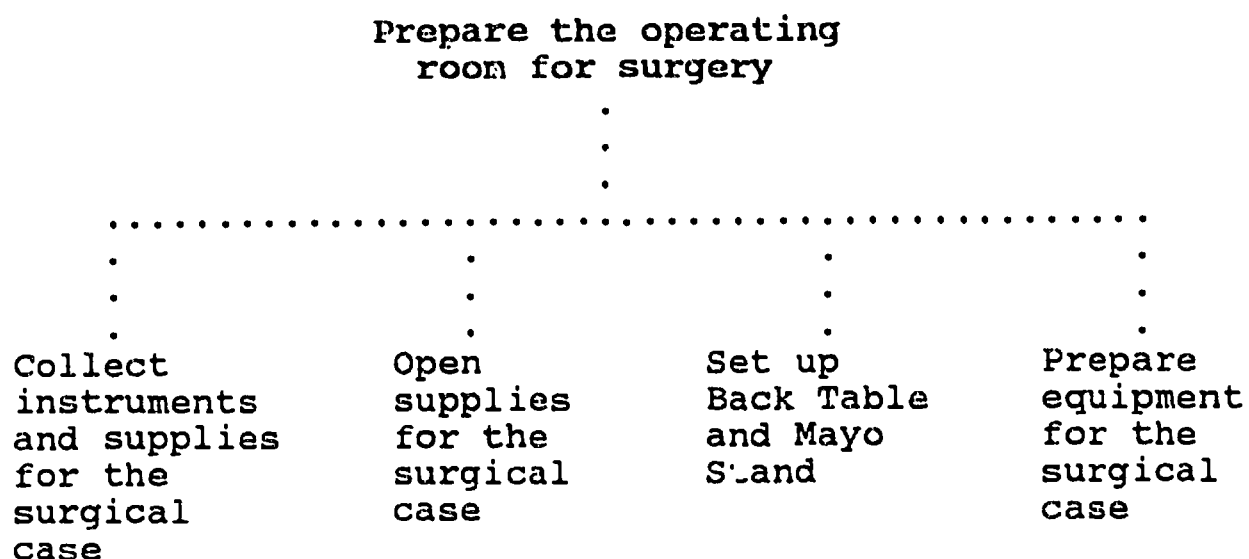


Figure 5. A breakdown of the task, Prepare the operating room for surgery

the job; to maintain a task orientation in their descriptions of the job; and to provide specific levels of detail in the descriptions.

### STEP 3: The Job Inventory Questionnaire

In order to obtain input from the entire surgical nursing staff, data from the final job-tree were used to develop a job inventory questionnaire. The job-tree approach proved to be very efficient in the development of the inventory. It was developed by listing each of the 161 sub-tasks in a questionnaire format. This allowed all 30 ORNs to indicate how often they performed each of these activities and the average amount of time it takes to perform them. The sub-tasks were grouped into five different sections corresponding to each of the five previously identified

sub-jobs. The responses of the surgical staff were tabulated for the 161 questionnaire items. These results were then incorporated into the job-tree as an indication of the importance of each job activity. The final job-tree, included a breakdown of the job of the ORN into sub-jobs, tasks and sub-tasks along with an estimate of the frequency in which sub-tasks are performed and the average amount of time that it takes to perform each sub-task.

### Discussion/Summary

The HJA provided a comprehensive look at the job of the ORN. It identified five distinct functions that are performed by nurses in the surgical department at this particular hospital. With the HJA we were able to describe the position of the ORN using a task orientation with several levels of detail. The value of this approach becomes apparent when the job analysis results are used to develop human resources programs.

One measure of the utility of a job analysis is in the ease with which the results are used in later programs. The breakdown provided by the HJA is well suited for designing performance appraisal systems, pre-employment selection instruments and training programs.

A performance appraisal instrument should focus on the major functions of the job. The job-tree approach clearly defines these major functions. In the present case study the major functions of the job are best represented at the "task" level of the job-tree. A performance appraisal instrument can be constructed to evaluate

these tasks. The sub-tasks can be used to develop items for measuring performance on each of the tasks. For example, to evaluate the performance of an ORN in the task of preparing the operating room for surgery, several items could be generated. One such item would evaluate how effective a nurse is at collecting all of the necessary instruments and supplies for surgical cases. Other items could be developed in the same manner.

The HJA is equally well suited for developing pre-employment selection instruments. Selection interviews could be developed by structuring questions which evaluate the applicant's ability or past experience in performing the major functions of the job. Questions for the interview could be constructed using the same procedure in which performance appraisal items are developed. A job candidate's experience or knowledge of setting up for surgical cases could be evaluated by asking the candidate to describe the procedure. The interviewer could refer to the sub-task items as a standard for evaluating the candidate's responses.

The job-tree is also effective for developing training programs. The five sub-jobs which were identified in the case study provide logical topic areas for training. A typical training program can be developed which consists of a sequence of training modules. The tasks that are listed below each sub-job on the job-tree clearly define the domain of the possible training modules.

In summary, we feel that the HJA is an effective technique, capable of producing a thorough description of the job. The value

of the HJA lies in the fact that the results of the analysis are well organized, allowing for ease in the application of the results to human resources programs. The HJA is a simple procedure. It provides a clear strategy for structuring the job analysis interview. The fact that it follows a logical sequence of steps from global to specific descriptions of the job makes the HJA attractive to those individuals who conduct the job analysis interview as well as those who are asked to participate in them.



## REFERENCES

- Arvey, R. D. & Faley, R. H. (1988). Fairness in Selecting Employees. Reading, Mass.: Addison-Wesley Publishing Company.
- Cascio, W. F. (1982). Applied Psychology in Personnel Management. Reston, Va.: Reston Publishing Company.
- Gatewood, R. D. & Field, H. S. (1987). Human Resource Selection. New York, N.Y.: Dryden Press.
- Rendero, T. (1981). Consensus. Personnel 58: 4-12.

TOWARD A GENERIC JOB ANALYSIS SYSTEM

JAI GHORPADE, Ph. D.  
Professor of Management  
College of Business Administration  
San Diego State University  
San Diego, California 92182-0096  
(619) 461-0396

Proceedings of the 14th Annual Conference of the  
INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION, ASSESSMENT COUNCIL  
San Diego, California  
June 24-28, 1990

Please address all correspondence relating to this paper to:

Professor Jai Ghorpade, Ph.D.  
Department of Management  
San Diego State University  
San Diego, CA 92182-0096

June, 1990

## TOWARD A GENERIC JOB ANALYSIS SYSTEM

A case is made for the development of a generic job analysis system that draws its elements from the properties of jobs viewed as organizational sub-units. To facilitate this development, ten properties of jobs that need to be taken into account in attempts to analyze them are presented. Criteria that a generic job analysis system will have to meet are specified. A progress report is presented of the author's attempt to develop such a system.

Even though job analysis has occupied a prominent place in the theoretical writings of human resource management since the beginning of the factory system (Uhrbrock, 1922; Ling, 1965), the concept languished in the background of industrial practice for many years. In recent years, however, job analysis has been brought to the forefront and occupies a prominent place in the theory and practice of human resource management. This development can be traced to the growing complexity of modern organizations as well as to legal requirements which mandate systematic gathering and analysis of data to be used in decisions relating to people (Ghorpade, 1988).

The growing popularity of job analysis has been accompanied by the development of an array of systems (1) of job analysis (Gael, 1988, Vo. 2). Unfortunately, the growth of these systems has been a mixed blessing. An obvious benefit of this development is that job analysts now have a wide choice in regard to method. But the existence of choice in regard to method adds positively to an undertaking only if the competing methods are alternative modes of performing the same activity. In fact, that is not the

- 
1. In the job analysis literature, the term system is used interchangeably with the following terms: technique, method and procedure. We prefer the term system as it is broadest in conception and the most developed conceptually.

case. Most of the existing systems of job analysis have little in common in regard to objectives, inputs, procedures and outputs. Furthermore, ratings by experienced job analysts have yielded highly significant differences in regard to their relative utility when judged against common purposes (Levine, Ash, Hall and Sistrunk, 1983).

This paper seeks to begin the task of assembling a job analysis system that draws its substance from the generic features of jobs viewed as organizational sub-units. The immediate objectives of this paper are: (1) to spell out the properties of jobs that need to be accommodated within any comprehensive system that is designed to study them; (2) to suggest some criteria that a generic system of job analysis needs to meet; and (3) to show how such a system can serve managerial functions in the handling of problems endemic to the employment relationship.

#### WHAT IS A JOB?

The literature on job analysis is now immense (Gael, 1988, Vols 1, 2). However, not many have paused to take systematic note of the properties of the social entity that they have sought to study. We attempt here to capture properties of jobs that need to be taken into account in attempts to design a system for studying them. Figure 1 presents alternative views of a job viewed from four perspectives, and serves as a guide for this discussion.

#### Location of Jobs

We begin by noting that jobs are found in organizations (Figure 1A). This claim may be obvious, but it is a critical starting point. Recognition of this fact closes some doors and opens others. Immediately, we can rule out any discussion of jobs as independent entities, with intrinsic

properties that are independent of the wider realities in which they are found. Acknowledging this connection, on the other hand, highlights the dynamic interdependencies that exist among jobs, people, organizations, and the wider community in which the action takes place leading to the following generalizations:

- (1) As living parts of organizations, jobs share all the properties of organizations; in fact, jobs are microcosms of organizations and hence it is both proper and fruitful when analyzing them to use the concepts that have been used to study organizations.
- (2) Organizations themselves are parts of larger systems - communities, regions and societies. This being the case, gaining membership within an organization provides the individual with a place within the wider system. Jobs thus connect people to organizations and to the wider entity in which the organization is located. This connection plays an important part in shaping and defining the identity of the individual.
- (3) Being social entities, work within organizations gets done through the culture of the organization or beliefs held by the members about the world in general and how it works, and the values that guide their notions of right and wrong (Sathe, 1985).

#### Jobs as Open Systems

Organizations are open systems, and hence jobs can also be viewed as such. Acknowledging the open system properties of organizations yields the following insights:

- (4) Jobs survive through dynamic transactions with their environments, which happen to be organizations; the relationship is mutually reciprocal with survival of a job hinging on the ability of the job-holder to deliver the outputs essential for organizational operations.
- (5) Jobs are created and sustained by organizations, but they are subject to influence from forces outside the organization. A prime source of this influence are the multiple memberships (worker, citizen, family member) that job holders simultaneously retain.

#### Duality of Jobs: Functional Specialization and Hierarchy

The sub-units of organizations tend to be differentiated in two principal ways: function and hierarchy. The organization chart is thus simultaneously a blue print for action as well as a career ladder. To

occupy a job is to perform a specialized activity within a hierarchically ordered plan of action. Recognition of this duality yields the following generalizations:

- (6) Jobs are the building blocks of organizations; they constitute the primary units of work groups, department and other sub-systems.
- (7) Jobs are the vehicles of specialization in the performance of differentiated units of work.
- (8) Jobs provide a place for the individual within the hierarchy of the organization; the role played by an individual within an organization usually signifies the status of that individual within the organization as well as the wider community in which the organization is located.

#### Stakeholders of Jobs

Organizations are sustained through transactions among multiple stakeholders (Carroll, 1989). Jobs, being microcosms of organizations, have their own set of stakeholders. Figure 1B shows the principal stakeholders commonly associated with jobs. Recognition of the relevance of these interactions provides the following insights for job analysts:

- (9) Jobs are sustained through contractual arrangements and understandings among stakeholders. This is the political side of jobs and needs to be reckoned with in trying to understand how jobs come into being, how they are sustained, and incorporated into the fabric of the organization (Keeley, 1988)
- (10) The exchange that takes place with reference to jobs is of two types: (1) exchanges of materials, information, and other substantive goods that are essential to the fulfillment of the output expectations of the job, and (2) political exchanges among the stakeholders relating to the distributive issues surrounding job success or failure.

#### CRITERIA FOR A SYSTEM OF JOB ANALYSIS

Research and commentary relating to the workings of job analysis systems has brought to the fore a set of criteria that such systems need to meet in order to survive the competitive and legal challenges that face them in today's world (Gatewood and Feild, 1990). In this part, we briefly identify these and show their relevance in the design of job analysis systems.



Four sets of criteria are presented: general social system; operational; practicality; and effectiveness.

#### General Social System Criteria

Since the unit of analysis of job analysis systems is a social reality, it is not unreasonable that they be held to a set of criteria that are relevant to assessing all contrived social systems. The following criteria are derived from some rudimentary ideas advanced by E. Wight Bakke in this regard (Bakke, 1959):

- (1) Consistency with reality: The conceptual framework comprising the system should be recognizable by those associated with the job as being the actual sort of thing to which they are related.
- (2) Comprehensiveness: In order to qualify as a job analysis system, the system needs to address all the significant attributes of jobs viewed as social units. Systems that deal with parts of jobs should be labelled as such. Failure to do so can give rise to questions relating to truth in labelling.
- (3) Clarity of linkages: Two types of linkages need to be specified: (1) cycle of activities - linkages among input, process, and output variables of the system; and (2) organizational uses - linkages among the outputs of the system to the uses to be made of them in managerial decision-making.

#### Operational Criteria

This set of criteria deals with the internal composition of the system and its readiness for use.

- (4) Standardization: System terminology should be both internally consistent and in conformance with industrial practice.
- (5) Ready-to-Use: Extent to which the system is tested, refined and ready to be used.
- (6) Construct validity: Does the system do what it is supposed to do? Does it do so with accuracy?
- (7) Reliability: Does the system yield consistent results?

#### Practicality Criteria

This set of criteria is not mandatory in the sense that failure to meet them can reflect badly on the worth of the system. They are rather

descriptive criteria which enable the user to assess the practical feasibility of using the system within a particular context.

- (8) Off-the-Shelf Availability: Can it be used as is or does it need adapting and tailoring before it can be used?
- (9) Training Time: Scope and amount of training required before the analyst is able to use the system correctly. Very closely related to the emerging concept of "user friendly" in the field of information systems.
- (10) Versatility: Number of uses to which system can be put in human resource management and other managerial concerns.
- (11) Adaptability: The extent to which the system can be modified and tailored to suit the demands of the job situation.
- (12) Sample Size: How many respondents or sources of information does the system require in order to ensure valid results?
- (13) Costs: Developmental costs, if any, plus the cost of materials, clerical and other operational expenses.

#### Effectiveness Criteria

This set of criteria deals with the quality of results provided by the system with reference to the uses and desired outcomes which led to the adoption of the system.

- (14) Relevance of output (content validity): Extent to which the outputs produced by the system are relevant to the purposes to be served.
- (15) Utility: Extent to which the system yields high quality results with reference to the functions for which it is being used.
- (16) Side Effects: Positive and negative legal, ethical and organizational effects of using the system.

#### GENERIC JOB ANALYSIS SYSTEM (GJAS):

##### A PROGRESS REPORT

Guided by the preconceptions about jobs and system criteria noted above, the author has been attempting to assemble a system of job analysis that conforms to those preconceptions. Testing of the system thus far has been through field projects completed by students in courses in human resource management. Even though all the pieces of the system are not fully in

place, experience gained from the field projects makes it possible to provide a skeletal outline of the significant components of the system.

### System Components and Interrelations

The components of the GJAS system and their interrelations are shown in Figure 2. The primary output of the GJAS system is a comprehensive job description. This output is used in two principal ways: (1) to derive standard information outputs typically used in human resource management (labelled as derivatives of job descriptions in Figure 2); and (2) as a general source of information about the job. Charts are under construction for linking elements within the job description and its derivatives to human resource management functions and other managerial activities (see, Ghorpade, 1988, Ch 6).

### Data Forms and Instruments

The output of the GJAS is expected to be mixed in regard to form. Most of the elements in the job description have been linked with structured instruments which yield quantitative ratings. Many of these are from existing job analysis systems, particularly the FJA system; others are original constructions. Items which cannot be quantified (e.g. mission statements, task descriptions) are linked to systematic description procedures such as the Sentence Analysis Technique contained in the Department of Labor system (U. S. Department of Labor, 1972), and to scales for assigning ratings to such items in regard to their relative importance, frequency of performance and other concerns (Ghorpade, 1988, Appendix G). In all cases, preference is given to instruments and procedures that are expandable and adaptable to suit the needs of particular organizational demands.

### Process of Job Analysis

The starting point of job analysis in the GJAS is the crystalization of the purposes to be served by the analysis in the performance of broader managerial functions such as human resource management, organization design and development, and human resource strategy. Once this is accomplished, the analyst works backwards in the chain of interrelations shown in Figure

#### 2. Example:

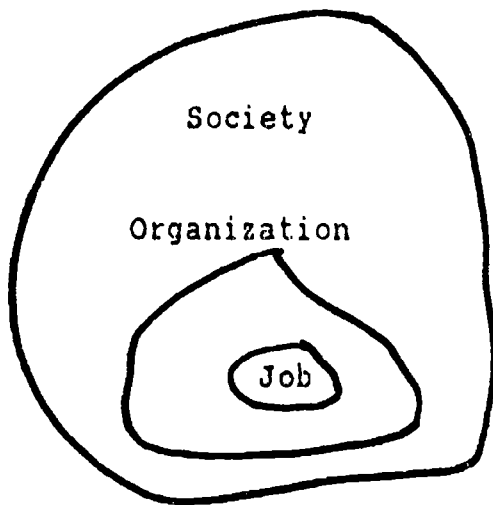
Suppose the organization is interested in improving its performance appraisal system. It would begin by examining existing statements of duties and responsibilities and performance criteria and standards. If these provide enough information for attaining the objectives of change, the analysis stops at that stage. On the other hand, if such statements do not exist or if the existing statements are inadequate, then the analyst proceeds to construct new ones or to reform existing ones. The GJAS manual (under construction) then directs the analyst to the items within the job description schedule that are most relevant for the attainment of these purposes.

### Statistical Programs

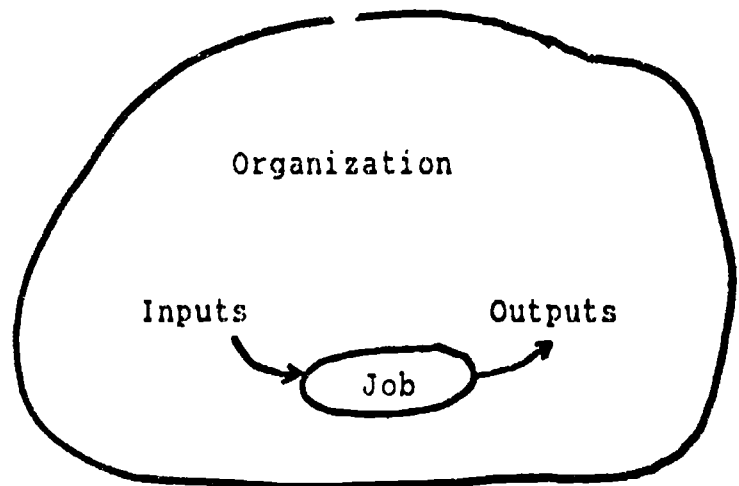
Eventually, statistical programs will have to be constructed to ease the task of data tabulation. However, it is accepted a priori that the process of job analysis is inherently a qualitative undertaking. Quantitative data will provide a basis for decision-making, but can never take the place of judgment. The thrust of the statistical program development activity will thus be on facilitating quantitative assessments of job elements that lend themselves to quantification and which can serve as basis for negotiation among the stakeholders of the job.

## LIST OF REFERENCES

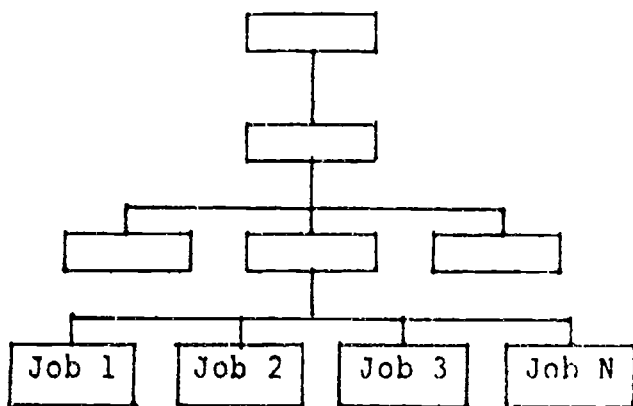
- Bakke, E. Wight (1959). Concept of the social organization. In M. Haire (Ed.), Modern organization theory (pp. 16-75). New York: John Wiley.
- Carroll, A. B. (1989). Business & society. Livermore, CA: South-Western.
- Gael, S. (Ed.). (1988). The job analysis handbook for business, government, and industry (Vols. 1-2). New York: John Wiley.
- Gatewood, R. D., & Feild, H. S. (1990). Human resource selection. Chicago, Illinois: Dryden Press.
- Ghorpade, J. V. (1988). Job analysis: A handbook for the human resource director. Englewood Cliffs, NJ: Prentice Hall.
- Keeley, M. (1988). A social contract theory of organizations. Notre Dame, Indiana: University of Notre Dame Press.
- Levine, E. L., Ash, R. A., Hall, H., & Sittrunk, F. (1983). Evaluation of job analysis methods by experienced job analysts. Academy of Management Journal, 26(2), 339-347.
- Ling, C. L. (1965). The management of personnel relations. Homewood, Illinois: Richard D. Irwin.
- Sathe, V. (1985). Culture and related corporate realities. Homewood, Illinois: Richard D. Irwin.
- Uhrbrock, R. S. (1922). The history of job analysis. Administration, 3, 163-168.
- U. S. Department of Labor (1972). Handbook for analyzing jobs. Washington, D.C.: U. S. Government Printing Office.



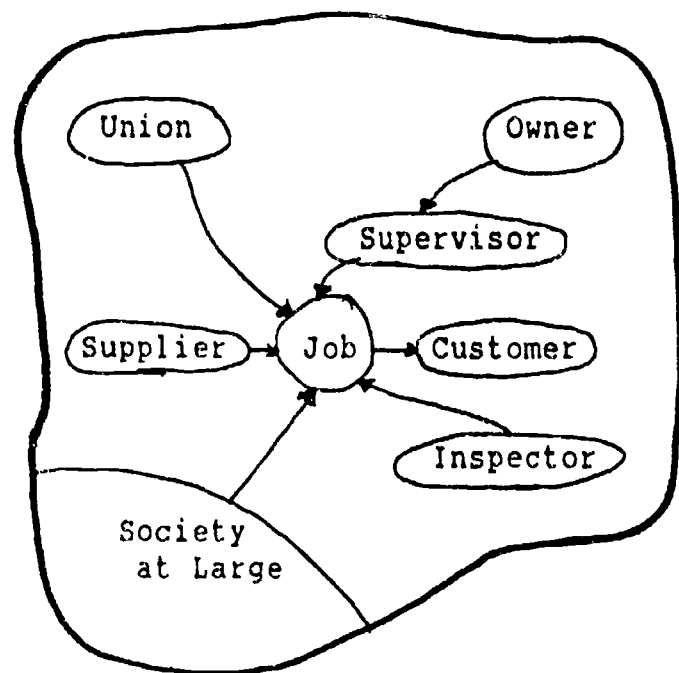
A. Job, organization & society



B. Job as an open system



C. Job as part of a bureaucracy



D. Stakeholders of jobs

Figure 1: Four faces of a job



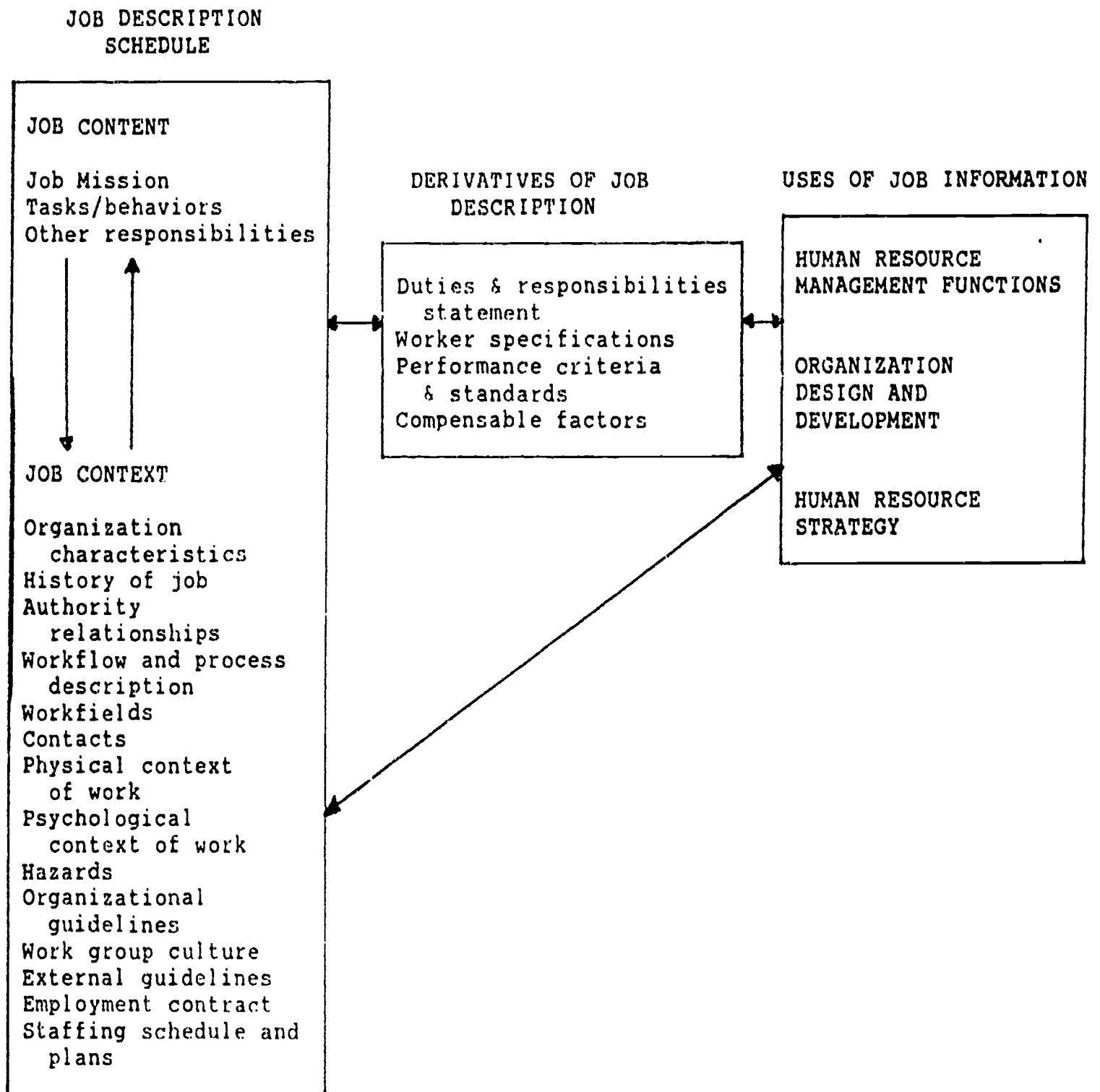


Figure 2: Interrelations among components of the GJAS

A Comparison of Job Information Descriptors  
for Classifying Jobs for Selection Purposes

Nelson Adrian

Los Angeles Unified School District  
315 East 21st St.  
Los Angeles, CA 90011

Paper presented at the annual conference of the International  
Personnel Management Association Assessment Council, San Diego,  
CA, June 28, 1990.

A Comparison of Job Information Descriptors  
for Classifying Jobs for Selection Purposes

Nelson Adrian

Personnel Selection Branch

Personnel Commission

Los Angeles Unified School District

Abstract

This paper summarizes a study which attempted to classify five positions, for selection purposes, where three sources of job information (task statements, KSAs, and general ability-oriented ratings) were analyzed separately by two data analysis methods. These methods involved assessing mean rating levels, and significant mean differences among the data. The results suggest that the type of job information collected, its specificity, and the method of analysis can influence the outcome of a job analysis.

## A Comparison of Job Information Descriptors For Classifying Jobs For Selection Purposes

There are numerous reasons to conduct research regarding the grouping of jobs. Job evaluation, job classification, test selection, and validity generalization are among the more common purposes. In order to classify jobs in meaningful ways there are two basic issues with which the classification needs to deal. The first is to determine what type of job analysis data or job descriptor information should be used to compare the various jobs. The second issue involves choosing an appropriate method for comparing these data to combine or categorize the jobs.

The most common types of job analysis data or job descriptor information include task-oriented approaches, in which each job is broken down into elemental units; worker-oriented job elements, involving generalized human behaviors required to do the work; abilities-oriented elements where underlying abilities and aptitudes required to do the work are studied; the critical incident approach where jobs are described in terms of critical incidents; and holistic judgements in which incumbents directly estimate the extent of various job attributes required.

There has been a substantial amount of research comparing different statistical models for analyzing job analysis data (see Lissitz, Mendoza, Huberty, and Markos, 1979; Cornelius, 1988). However, there are only two studies that have compared which type of job descriptor is best suited for a particular human resource purpose. Most of the studies addressing this topic have compared job analysis methods and these methods differ with respect to the job descriptor inherent to the method.

Cornelius, Carron, and Collins (1979) analyzed seven foreman jobs in a chemical processing plant using task-oriented, worker-oriented, and abilities-oriented job descriptors. Initially 373 task statements, written at the level of specificity outlined by Fine and Wiley (1971), were collected. There the Position Analysis Questionnaire (see McCormick, Jeanneret, and Mecham, 1972) was used to gather worker-oriented job information. The unit of analysis was generalized human behavior required to do the work. Finally, the abilities-oriented job descriptor information proposed by Fleishman (1972, 1975) was used. The unit of analysis was the underlying abilities and aptitudes. Jobs were studied in terms of the profiles of abilities required to do the work.

All three data sets were analyzed using a hierarchical clustering algorithm described by Ward (1963). Results for the task data indicated that either three or five clusters of foreman jobs could be identified depending upon the criterion of task overlap used. For the ability data, three clusters were specified. Only one cluster was found using the worker-oriented data. That is, essentially all of the jobs were found to be identical.

The analysis revealed different groupings regarding both the number of similar foreman jobs and which jobs were most similar. The authors concluded that, in at least some settings, various types of job analysis data may lead to different conclusions regarding job similarities.

Another study suggesting that the job analysis method and/or job descriptor is important in affecting the job analysis outcome was conducted by Levine, Ashe, Hall, and Sistrunk (1983). They were studying job analysis methods, but each method involved a different descriptor so the results may be, in part, due to the difference in descriptor rather than method. They administered a survey which gave descriptions of seven job analysis methods and asked questions concerning their effectiveness for each of eleven organizational purposes. Respondents were experienced users of job analyses and included individuals from universities, governmental agencies, private businesses and private consultants. etc. The job analysis methods included the critical incident technique (Flanagan, 1954); Position Analysis Questionnaire (McCormick et al., 1972), job element method (Primoff, 1975, cited in Levine et al, 1983), ability requirements scales (Fleishman, 1975), functional job analysis (Fine & Wiley, 1971), task inventory CODAP (Christal, 1974, cited in Levine et al, 1983), and threshold traits analysis (Lopez, Kesselman, and Lopez, 1981). For job classification purposes, the task inventory and functional job analysis (FJA) methods were rated highest, although the FJA was not rated significantly higher than the Position Analysis Questionnaire (PAQ) method. The critical incident technique was rated significantly lower than all other methods.

Levine, Ashe, and Bennett (1980) compared four different job analysis methods for developing examination plans in a civil service testing situation. The four methods included a worker-oriented descriptor (PAQ), a critical incident descriptor, a task-based descriptor and an abilities-oriented descriptor. The examination plans did not differ across the four methods. A panel of experts rated all plans as job-related.

Finally, Cornelius, Schmidt, and Carron (1984) compared, as the focus of their study, an elaborate task-based (activities element and abilities element) inventory procedure with simple job classification judgements by supervisors and incumbents. The data were analyzed using a Multiple Discriminant Analysis. Both the activities element and the classification judgements methods produced accurate classifications of jobs (96%) from 54 petroleum - petrochemical plants. Jobs were categorized as belonging to operations, maintenance, laboratory, or miscellaneous. However, it was also found that the activity elements portion of their questionnaire did a better job of classification (96%) than did an ability elements portion (80%) included in the questionnaire.

In conclusion, there is very little research directly comparing types of job information descriptors. Of the research that does exist three of four studies suggest that there are

meaningful and practical differences in job analysis results depending on the type of descriptor used. Of these, two of the three papers suggest to some degree that task based descriptors are better suited for job classification purposes than were worker or ability oriented descriptors.

The current study will compare task-oriented, worker-oriented, and ability-oriented job analysis descriptors, with the expectation that the more specific the type of descriptor the more sensitive it will be to differences among different classifications. Specifically, task-oriented descriptor items should be more sensitive than worker-oriented, which should be more sensitive than ability-oriented descriptor items.

### Methods

Classifications Studied. The study involved five separate classifications - Cafeteria Manager I through Cafeteria Manager IV (CM I - CM IV) and Satellite Kitchen Supervisor (SKS). The distinctions among the Cafeteria Manager I - IV positions are primarily based on the number of man-hours assigned to the cafeteria, which is a function of the number of students at the particular school. These positions are all located in elementary, junior high, or senior high schools belonging to a large school district. The Satellite Kitchen Supervisor is assigned to small schools which do not have a cafeteria. The SKSs supervise workers who heat and serve pre-packaged meals to students rather than preparing the meals in an on-site kitchen.

Sample. Six hundred and twenty-seven questionnaires were sent to all of the school districts current Cafeteria Managers, Satellite Kitchen Supervisors, and their supervisors. Two hundred and thirty one were returned (37%). Questionnaires were completed by 13 CM Is, 63 CM IIs, 4 CM IIIs, 44 CM IVs, and 43 SKSs from 106 elementary, 25 junior high, and 15 senior high schools. Only 6 of 26 supervisors returned questionnaires (each completing two questionnaires). The respondents included 5 Asians, 21 Blacks, 18 Hispanics, and 61 Whites. The remaining 126 did not identify their ethnicity. One hundred forty eight females and two males completed questionnaires and reported their gender. The average age of respondents providing their ages was 50.05 years.

Measures. A questionnaire which consisted of 43 task statements and 18 knowledge, skill, and ability (KSA) statements was developed from interviews conducted with incumbents and supervisors. A general abilities-oriented questionnaire consisting of 33 psychomotor abilities adapted from Fleishman (1975) was also included. Thus, there were task-oriented, worker-oriented, and general abilities-oriented sections of the questionnaire.



Tasks were rated on "Relative Time Spent" and on "Criticality" using a four point scale; KSAs were rated on "Expected at Entry" and "Criticality" also using a four point scale. The general abilities were rated on a five point behaviorally anchored scale.

An alpha coefficient was calculated for the task statement ratings to assess the reliability of these ratings. Task items relating to the same factor were grouped to run this analysis. This analysis was performed to provide an indication of at least whether respondents were completing the questionnaire in a careful, thoughtful manner. Alphas of .81 for Food Preparation; .54 for Sanitation, Hygiene, and Safety; .92 for Supervision; .74 for Food Ordering; .79 for Record Keeping; and .80 for Interpersonal Relations was found.

Classification Procedure. The three sources of job information were analyzed separately by two data analysis methods for comparison purposes. The first method involved assessing the mean ratings and the second method employed inferential statistics to assess mean differences.

The first method used to judge similarities and differences among the five positions involved looking directly at the mean ratings for each position. These ratings indicated which task, KSA, or ability was rated as part of the job. A mean rating of 1.0 or higher for the task Time Spent and Criticality ratings, and for the KSA Expected at Entry and Criticality ratings, and a mean rating of 1.5 or higher for the general ability ratings were used as criterion measures. Ratings at or above these levels indicated that it was a significant part of the job as described by the questionnaire scale anchors.

The second method involved performing a MANOVA on the ratings. This was followed by a series of ANOVAs to determine which tasks, KSAs, and general abilities had significant differences. Where significant differences were found Scheffe t-tests were used to determine among which positions the differences existed.

## Results

The two data analysis procedures failed to indicate any differences in the general ability ratings across the five classifications. All of the general abilities were rated as being a part of performing the job and the MANOVA did not indicate any differences (see Table 1). All five classifications could be grouped together based on the general ability ratings.

insert Table 1 about here

Assessing the mean KSA ratings suggested a similar conclusion. As indicated in Table 1 by the percentage of ratings in common among the five classifications, at least 94% of the KSAs were shared among all five positions. SKS had 2 KSA ratings which indicated that they were not part of the job and CM I had one. These related to supervision KSAs. All ratings for all of the other classifications were rated as part of the job.

Task ratings yielded the greatest number of differences. The mean ratings still indicated that all five classifications had at least 79% of the job duties in common (see Table 1).

The MANOVA indicated that significant differences exist among KSA ratings. Subsequent ANOVA and Scheffe tests indicated that the number of significant differences ( $p < .05$ ) ranged from 8 of 36 KSA ratings to 0 ratings. In other words, KSAs shared among the five classifications ranged from 78% to 100% (see Table 1).

The MANOVA also indicated that significant differences existed among task ratings. Subsequent ANOVA and Scheffe tests revealed that the number of significant differences ( $p < .05$ ) ranged from 36 of 84 task ratings to 0 of 84 ratings. Table 1 shows the percent of tasks shared among each position which ranged from 58% to 100%.

### Discussion

The results demonstrate that the type of job descriptor used in the job analysis questionnaire is important when classifying jobs for selection purposes, as different results were generated using different job descriptors. Based on the MANOVA analysis the general ability ratings made no distinctions whatever among the five classifications. The KSA ratings showed some differences but no less than 81% of the KSAs were shared among the classifications. Conversely, the task ratings yielded the most differences among classifications sharing as few as 58% of the job tasks to a statistically significant degree.

The results were also dependent upon the type of data analysis employed. Analysis of only the mean ratings suggested that no less than 79% of tasks were shared by all classifications, versus as few as 58% using the MANOVA analysis. Overall, the results suggest that these five classifications are quit similar, with the exception of SKS.

The exact point at which one should combine different classifications together for selection purposes is open to debate. However, classifications which are 79% similar can probably be grouped for selection purposes in a defensible manner, while classifications only 58% similar probably can not be grouped without making modifications to the examination.

These results emphasize the need to carefully consider the type of job descriptor used when analyzing a group of

classifications or positions for classification purposes. Further, the classification purpose itself must be considered. These results suggest that the more specific the job descriptor, the more likely it is that differences among classifications will be found. One might choose a job descriptor with a particular outcome in mind. It also appears from these results that different classification outcomes may occur based on the data analysis procedure employed. Thus, since it is not practical to use multiple data descriptors and multiple data analysis methods all of the time, future research needs to address under what circumstances various descriptors and analysis methods are most appropriate. Further, a scientific approach to job classification cannot allow one to try several job analysis methods in order to choose the one which suits one's own needs or desires the best.

## References

- Cornelius, E.T. Practical findings from job analysis research, in S. Gael (Ed). The job analysis handbook for business, industry, and government. New York, John Wiley & Sons, 1988.
- Cornelius, E.T., Carron, T.J., and Collins, M.N. Job analysis models and job classification. Personnel Psychology, 1979, 32, 693-708.
- Cornelius, E.T., Schmidt, F.L., and Carron, T.J. Job classifications approaches and the implementation of validity generalization results. Personnel Psychology, 1974, 37, 247-260.
- Fine, S.A. and Wiley, W.W. An introduction to functional job analysis, Washington, D.C.: The W.E. Upjohn Institute for Employment Research, 1971.
- Flanagan, J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.
- Fleishman, E.A. Toward a taxonomy of human performance. American Psychologist, 1975, 30, 1127-1149.
- Levine, E.L., Ash, R.A., and Bennett, N. Exploratory comparative study of four job analysis methods. Journal of Applied Psychology, 1980, 65, 524-535.
- Levine, E.L., Ash, R.A., Hall, H., and Sistrunk, F. Evaluation of job analysis methods by experienced job analysts. Academy of Management Journal, 1983, 26, 339-348.
- Lissitz, R.W., Mendoza, J.L., Huberty, C.J., and Markos, H.V. Some further ideas on a methodology for determining job similarities/differences. Personnel Psychology, 1979, 32, 517-528.
- Lopez, F.M., Kesselman, G.A., and Lopez, F.E. An empirical test of a trait-oriented job analysis technique. Personnel Psychology, 1981, 34, 479-502.
- McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 1972, 56, 347-368.
- Ward, J.H. Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, 1963, 58, 236-244.

TABLE 1

## PERCENT OF CLASSIFICATIONS RATED AS SIMILAR BY METHOD

Comparison Classifications

<u>DESCR</u> <u>IPTOR</u>	<u>CMI/</u> <u>CMII</u>	<u>CMI/</u> <u>CMIII</u>	<u>CMI/</u> <u>CMIV</u>	<u>CMII/</u> <u>CMIII</u>	<u>CMII/</u> <u>CMIV</u>	<u>CMIII/</u> <u>CMIV</u>	<u>SKS/</u> <u>CMII</u>	<u>SKS/</u> <u>CMIII</u>	<u>SKS/</u> <u>CMIV</u>	<u>SKS/</u> <u>CMIV</u>
RATINGS										
Tasks	96%	91%	88%	94%	92%	95%	88%	85%	79%	79%
KSAs	97%	97%	97%	100%	100%	100%	97%	94%	94%	94%
Ability	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
MANOVA										
Tasks	100%	93%	86%	95%	84%	99%	87%	70%	83%	58%
KSAs	94%	97%	97%	100%	100%	100%	100%	81%	97%	78%
Ability	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%