DOCUMENT RESUME

ED 337 484 TM 017 304

TITLE	Proceedings of the 1988 IPMAAC Conference on Personnel Assessment (12th, Las Vegas, Nevada, June 19-23, 1988).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	Jun 88
NOTE	228p.; Pages 175-179 have extremely small print which may not reproduce clearly.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	Assessment Centers (Personnel); Computer Assisted Testing; *Evaluation Methods; Job Analysis; *Job Performance; *Occupational Tests; Personality Assessment; *Personnel Evaluation; Personnel Management; Personnel Selection; Psychological Testing; *Public Sector; Screening Tests; *Test Use
IDENTIFIERS	International Personnel Management Association

ABSTRACT

Author-generated summaries/outlines of papers presented at the annual conference of the International Personnel Management Association Assessment Council (IPMAAC) in 1988 are provided. The "Presidential Address" is by N. E. Abrams. The keynote address is "Is There a Future for Intelligence?" by R. Thorndike. Summaries of 53 papers on the following selected topics are provided: assessing productivity; use of video technology in testing; supervision; selection criterion; screening direct care workers for child abuse potential; screening models for psychological testing; examination security; an interactive oral exam for juvenile correction workers; multiple-choice questions that malfunction; bias and test-wiseness; the promotability index; tests for selecting 911 telephone operators; direct versus indirect writing assessment; videotaped work incident simulations in police and fire assessment centers; predicting job performance of mentally retarded persons; employment of the disabled; a multi-purpose job information system; test security, applicant rights, and the candidate review process; selection of police managers; job satisfaction in the Federal workforce; paper and pencil measures versus assessment centers in police selection; validation of physical performance tests; use of departmental ratings of promotability; personality testing; criterion-related validation using two-way validity generalization; application of Angoff in passing point setting for a situational interview; using the Social Skills Inventory in personnel assessment; the Worker Characteristics Inventory--a methodology for assessing personality during job analysis; refining a self-rating selection instrument--correction of self bias; computerized testing made practical; guides in the design of simulation exercises; and the development of job-related medical standards/guidelines for selection of applicants and evaluation of incumbent personnel. An author index is provided. (SLD)

IPMA Assessment Council

Proceedings of the 1988 IPMAAC Conference on Personnel Assessment

June 19-23, 1988

Las Vegas, Nevada

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

PROCEEDINGS OF THE TWELFTH ANNUAL
INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION
ASSESSMENT COUNCIL, 1988

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

The PROCEEDINGS essentially summarize presentations from information available to the program chair. All presenters, with the exception of invited and keynote speakers, were required to limit the published version of their papers to approximately 4 pages. Hence, some presenters were able to include most of their oral presentations while others opted to provide only topical outlines. Papers were published in the condition received, without editing. A few authors did not provide a written version of their presentations, and hence their presentations are not included in this volume.

While many tables and statistical data are included, others had to be excluded because of length. However, bibliographies are included if they were available. Persons interested in additional information regarding a presentation should contact the author(s) directly in order to determine if a more complete paper is available.

PREPARED UNDER THE GENERAL DIRECTION OF:

Wendy J. Steinberg
New York State Dept. of Civil Service
Albany, New York 12239

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

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

- providing opportunities for professional development;
- defining appropriate assessment standards and methodology;
- increasing the involvement of assessment specialists in determining professional standards and practices;
- improving practices to assure equal employment opportunity;
- coordinating assessment improvement efforts.

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

Published and distributed by the International Personnel Management Association Headquarters:

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

Refer any questions to the IPMA director of Assessment Services.

Table of Contents

(Titles may be abbreviated to conserve space; first authors only listed)

PRESIDENTIAL ADDRESS

"LOOKING TO THE FUTURE:
IPMAAC AND APPLIED PERSONNEL ASSESSMENT"

Nancy E. Abrams, Ph.D.
IPMAAC President

The year 1987-1988 has been a critical one for IPMA as an organization. For many years we have dealt with the issues associated with formation of a new organization, especially when the new organization is part of an already existing organization. At last year's meeting we reached a cross-road: would we be able to resolve our continuing organizational conflicts with IPMA or did we need to strike out on our own. The problems stemmed from financial and autonomy disputes with IPMA.

At this point in time, I believe that we are well on our way to resolving the issues. Naturally work on our part as well as that of IPMA is needed to keep things on the right track, but our mutual understanding and improved communication should go a long way to resolve our long-standing problems.

Given this situation, I believe that it is time for us to think about where we go from here. I thought that it was a good time for us to go back and think about why IPMAAC was formed twelve years ago. Why is there an IPMAAC? I'm sure that we all have our own personal answers to that question but perhaps it is time to look at the broader answer. The bylaws of IPMAAC lists seven purposes for our organization. I think that now is the time for us to review these purposes, determine if they are still relevant, what we have been doing to meet these purposes and what additional things we need to do the meet them.

"1. To support the general purposes and methods of the International Personnel Management Association--United States; in particular, to serve as a resource of professional expertise on technical policy matters."

This still seems to be an appropriate purpose for our organization, especially since we have decided to remain a part of IPMA.

I believe that we need to become more of a focal point for IPMA to use our professional expertise on technical policy matters. We have been represented in groups formed to comment on APA standards and Federal Uniform Guidelines.

We should be used as a resource in IPMA to assist Assessment Services. These services clearly fall within the area of expertise of IPMAAC members. Perhaps an advisory group from

IPMAAC could assist and advise on technical issues. IPMA has a test sharing service, we should provide input on how that can be done in the most technically sound way.

"2. To encourage and give direction to public personnel assessment maintenance and improvement efforts in fields such as training evaluation, job analysis, and organizational effectiveness."

Again this purpose continues to be an appropriate one for us, especially as personnel assessment organizations are threatened by finding cuts and other attacks.

Again I believe we can do more. The Research Advisory Committee has existed as a resource for those with technical questions or problems to help provide direction. This resource has been used by few IPMAAC members. The Committee is planning to develop a directory of persons with expertise in particular areas so that appropriate expertise can be identified.

Our conference and our publications also provide resources toward this purpose but we need to be viewed more as a resource not only by individual members but also the organizations for which they work. We should be thought of as a place to go when difficult technical questions arise. We should be actively trying to define good practice. We should provide support to employers trying to improve their practices or when they are being threatened.

Perhaps question and answer sessions at the conference on specific topics might be held. Perhaps we should consider going beyond our monographs, developing how to do it or procedural manuals as a series?

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

Especially in times of scarce resources, which seem to occur frequently, especially in the public sector, this seems a very valuable purpose.

We have been quite successful in the information exchange part of this purpose. PASS and ACN, in addition to the conference, are all vehicles designed to facilitate information exchange. Naturally, there is room for improvement. The more we can broaden our base of information exchange the more useful the exchange will be.

On the other side of this purpose, we have had little success. When the IPA grants stopped, most of this stopped too. WRIB and the efforts of some of the consortia have continued cooperative or poolings of resources. IPMAAC has done little of this. Can

cooperative efforts be done on a national bases? Is this too unrealistic a goal for us? I don't know.

The IPMAAC Selection Specialist Job Analysis proved that a nationwide study such as that could be done even without funding. Perhaps as we develop products from that effort, we can use the funds to support other large scale efforts.

"4. To define professional standards for public personnel assessment."

I am not sure that we should confine our purpose to the public sector. Each year we seem to draw more and more participants from the private sector. Perhaps we should say "applied personnel assessment." As a person who works in both the public and private sector, I gain information from IPMAAC useful in both spheres of work. For me, what sets IPMAAC apart as a valuable resource is that we deal in the real world rather than in theory. We are looking to solve real problems.

Should we be defining standards? Should there be an IPMAAC Standards apart from the APA standards? Perhaps a better solution would be a series of issue papers on controversial topics, perhaps even as part of the monograph series: Issues of particular relevance to us such as pass point setting, ranking, content validation, job analysis, etc.

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

This is clearly a purpose on which we have expended a great amount of effort. We have 2 three-day workshops which are offered on a regular basis (T & E and Examination Planning) and one on statistics to appear next year. We have offered preconference workshops on a variety of topics at this and the IPMA and IPMA regional conference. We will be discussing this particular area with the consortia to determine ways we may be able to work more closely on this.

Should we be providing more input on formal training? One goal of the Selection Specialist Job Analysis was to define training needs for various activities and communicate this information to colleges and universities so that they might consider developing programs to meet these needs. I believe that this is still a useful endeavor.

"6. To contribute to the formation of public policy relating to public personnel assessment."

Again this seems to be an appropriate role for our organization. However, we have not been very active in this arena. Through

IPMA we have been ready to comment on draft revisions of the Uniform Guidelines on Employee Selection Procedures. There have been no drafts to date to review.

Perhaps we should be taking a more active role in commenting on proposed legislation or at least notifying our members of such proposals? Perhaps we should intervene in relevant court cases, but this is very costly. At least, again we can see our members know when decisions are handed down or even issues involved in currently being tried or just heard cases.

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

Again this seems to be a very appropriate role for us, but a difficult one to operationalize. We have very slight progress in this area. We have been invited to speak before the Association of State Legislators.

What else should we be doing? I'm not sure but I am sure some of you may have some ideas.

After reviewing this list, I believe that the reasons IPMAAC was formed 12 years ago are as fresh and perhaps more relevant to us today as they were in 1976.

There is still a great need for a professional association of assessment professionals. In my opinion, what is needed is to greatly expand our scope, vision, and influence. I do not plan to retire. I look forward to working to expand the scope of IPMAAC and invite you all to do the same.

IPMAAC KEYNOTE ADDRESS

IS THERE A FUTURE FOR INTELLIGENCE?

Dr. Robert Thorndike, Professor Emeritus
Teachers College, Columbia University

For 70 years now, man and boy, I have been involved with ability testing. One of my early memories is of being dragged from bed one evening, sleepy and protesting, to serve as the guinea pig in a demonstration to a graduate student group at Teachers College of what I have subsequently come to recognize as the then quite new Standford-Binet Intelligence Test.

After that, I took most of the tests that were given in school or that I found kicking around my father's study, so that I became one of the most test-wise youngsters in that test-naive era. By the time I got to college I was able to bust the top off the guidance test given during freshman orientation week -- with the result that I became a chronic under-achiever. My college record could never quite come up to that test score.

After a slight side-excursion, as a graduate student, into studying the intelligence, if any, of chickens and rats, I settled down to do research on ability tests, to teach about ability tests, to write books about ability tests, and, over the past 30-odd years, to produce ability tests. There is certain poetic justice that my final enterprise has been the preparation of a new version of that same Standford-Binet that I first took seventy years ago.

Eighty plus years ago Alfred Binet was the first to produce what might be called an intelligence test. Moved by the need to differentiate between those who could not profit from the instruction in Parisian schools as they were then organized and those who would not, he assembled an assortment of tasks, graded in difficulty, that could be presented in a standard way to children, to determine at what cognitive level they were functioning. The tasks called for memory, judgment, comprehension and reasoning. Each was tried out on school and institutional groups of various ages to make sure that it did differentiate between the younger and the older children and between those in regular classes and those in institutions for the mentally retarded. Only tasks that met this standard were retained. The final product was well received, especially in the US, and was quickly adapted to the American scene, most notably in the Stanford-Binet authored by Lewis Terman.

Binet never paid too much attention to the theoretical basis for his test. He believed that an effective test should be based on tasks calling for relatively complex mental functions. But within this framework, his approach was primarily pragmatic, assembling a considerable variety of tasks that could provide a series of graded difficulty but of no one form. From the mixture, he believed something of practical utility would emerge -- and, indeed, he was correct.

At essentially the same time that Binet was assembling his test in France, the Englishman Charles Spearman was developing a statistical and theoretical rationale that provided a logical basis for Binet's hodge-podge approach. studying a considerable array of measures of ability and academic performance, Spearman found that each of them showed positive correlations with all the others, correlations that appeared to fall into a simple and orderly pattern. Spearman developed statistical procedures for analyzing that pattern which were the forerunners of modern factor analysis. He thought that the pattern of relationships could be accounted for by one single common factor running through all of the different measures, and he labelled it g to signify its generality. Some test tasks drew more heavily on g and some less, but this was the one thing that they all had in common. In addition to g he believed that each task depended upon some specific ability factor unique to that task. A reasonable approximation to a measure of g emerged from pooling the diverse assortment of tasks that Binet had included in his scale, and this gave coherence and meaning to the resulting score.

As time went on it became clear that a single general factor didn't tell the whole story of human cognitive ability. With the development of a wider range of tests, and of more sophisticated methods of correlational analysis, it became clear that certain tests had more in common with one another than could be accounted for simply by their loading with g. Additional ability factors were required. Techniques of multiple factor analysis, developed in large measure by L.L. Thurstone at the University of Chicago, were applied to tease out a number of distinct "Primary Mental Abilities" form comprehensive test batteries. In Thurstone's work each test was thought to depend on one or more (but preferably only one) of these primary mental abilities, and each of the primaries was thought to appear in only a fraction of the tests. Some of the primaries that were identified were such factors as Verbal, Numerical, Spatial, Inductive Reasoning, Deductive Reasoning and Memory. From the 1930's on, factor analytic studies led to a proliferation of factors until in Guilford's 1967 Structure of Intellect the number had been expanded to 120 in a neat, but somewhat unrealistic, 3-dimensional model.

Many tests have been produced in part to predict success in different jobs. Job analysis suggested that different jobs called for different abilities, and tests were concocted to appraise these different abilities. Studies multiplied in which a group in some occupation-- unfortunately, usually a small group-- took a battery of tests to see which ones would yield a prediction of measures of success in that job. But there were some recurring themes in the results, with measures of mechanical comprehension, clerical speed and accuracy, spatical perception, verbal and numerical ability, as well as general reasoning and problem solving, showing up in different settings as having promise as predictors.

Aptitude test batteries designed to appraise a number of different abilities reached their peak during and in the decade or two following World War II. In the Air Force we administered an Aircrew Classification Battery to well over a million men to sort candidates into those to be sent to pilot training, to navigator training or to bombardier training, and to weed out the also-rans. Studies of the validity of the tests in the battery were carried out on literally thousands of candidates, and test weighing procedures progressively refined. It was only with groups of this size that weighing schemes showed a reasonable degree of stability from one sample to the next.

During the same period, the U.S. Employment Service developed the GATB -- the General Aptitude Test Battery -- for civilian job counseling and guidance, and gathered validation data on over 400 different jobs. The accumulation of test results had led, on the one hand to a doctrine of job specificity in prediction, and on the other to the development of these comprehensive multiple ability batteries to cover the abilities that appeared to recur in different settings. The doctrine of job and situational specificity was the Gospel in personnel research and became engraved in stone in the EEOC regulations: ability tests must be specifically validated for each situation where they are used to make personnel selection or classification decisions.

In the enthusiasm for identifying and measuring specific ability factors, the role and even the existence of any general cognitive ability was often lost sight of. But it was still true, as Spearmen had observed much earlier, that the different tests in these batteries all tended to show positive correlations with one another. And though it was possible to account for these correlations by teasing out a number of separate factors, no one of which appeared in all of the tests, this could only be accomplished by resorting to factors which were themselves correlated. The general ability was still there but it had been buried in this correlation among the factors themselves and largely ignored in much of the literature on personnel testing.

Factor analysis does not explain the relationships among an extended set of tests. It serves only to provide a simplified and clarified description of those relationships. And there is no single correct description. There are an unlimited set of descriptions that are mathematically equivalent, and from that point of view, equally correct. The choice must be the one that is most helpful in clarifying the underlying structure of the set of variables or in arriving at useful relationships between tests variables and the events of the "real world".

To illustrate, I have taken data from the ten subtests of our Cognitive Ability Tests, Form 3. The subtests were designed to assess three distinct ability factors -- verbal, quantitative and visuo/spatial -- with the recognition that all of the tests also assess general cognitive ability. The correlation among these ten subtests have been factor analyzed by standard procedures, and the results are shown in Table 1. In a table of factor loadings, the size of the loading indicates how completely the test scores for a given test can be accounted for by that factor.

This table shows two mathematically equivalent representations of the observed correlations. The two display identically the same facts, and either can be derived from the other. Analysis A accounts for the correlations with no general factor. Here the large factor loadings indicate that the first four tests cluster together on the first factor, the next three tests have large loadings primarily on the second factor, and the last three tests on the third. But all of the tests have appreciable loadings on all of the factors. No sub-test is a pure measure of just one of the three factors. Analysis B proceeds differently, first extracting a general factor that includes whatever is common to all ten tests. Then the other factors pick up the more limited relationships that still remain between sub-tests designed to measure a single factor.

I believe that Analysis B gives a clearer portrayal of what is going on in these ten tests, for it makes it clear that there is a common factor running through all of the tests. This general factor is actually predominate in each one of the tests -- each test has its largest loading on the general factor. This analysis shows that the specific factors are real, but of relatively minor influence on the test scores. The differential information that we get from arranging the ten subtests into the three test scores -- Verbal, Quantitative and Nonverbal -- is pretty limited, and they all share the bulk of the information that each can provide.

Now let's look at Table 2 for some facts about published tests into which these ten subtests have been combined. Section A shows the test-retest reliabilities over roughly a six month

period, together with the correlations among the three tests. The reliabilities are reasonably satisfactory, and the intercorrelations are lower than the reliabilities, but not as much lower as one might like.

Combining the three tests appropriately weighted, produces the best estimate of g, the common factor that they share. Section B shows the weight to use for each in forming a composite and gives the reliability of that composite. Clearly, pooling the three tests provides a very dependable estimate of general cognitive functioning.

Section C demonstrates how much confidence we can have in the differences between pairs of tests. When we look at differences, we largely remove the effect of g because this is common to both tests. The variation attributable to genuine difference is larger than that resulting from measurement error, but only slightly so. In contrast to the highly reliable estimate of general ability that can be obtained from pooling the three tests, the differences that appear between verbal and quantitative, verbal and non-verbal, or quantitative and non-verbal are distressingly unstable.

Another way to look at the picture is to ask what fraction of our ability to predict performance can be accounted for by g, and what part is dependent on abilities peculiar to each specific training program or job?

There have been dozens of studies relating scores on batteries of tests to appraisals of success in different work settings. But most of these have been on small groups and have not been replicated. Results vary widely from one study to another, especially where complex weighing of the tests in a battery is involved. What is essential is to determine how well a particular selection procedure holds up when applied in a new sample of cases -- a procedure called cross-validation.

To illustrate, I have located data sets from two useful studies and done double cross-validation on each. The procedure involves determining an optimal set of test weights for sample A and applying those weights to sample B. Similarly, the optimal weights for sample B are applied to sample A. The validity in the crossed sample is compared with the validity of a general g factor estimated in a uniform way from the same battery and applied to both samples. The results are summarized in Table 3. In the first data set, validities were available for an Army battery of ten tests as predictors of end-of-course grades in 35 Army training schools ranging from Radar Repairman to Stenographer to Cook. Validities had been reported for two successive classes, so multiple regression weights could be determined on one class and then applied to the other. Classes typically enrolled about 250 men. The regression weighted composites were

compared with an estimate of g general ability applied uniformly to the data for both classes in each of the 35 schools. Results appear in the first column of the table.

With these groups, validity on the cross-validation sample was no more than 88% of that in the original group. However, in spite of the diversity of training programs represented in the data set, the g factor accounted for about 91% as much validity as the cross-validated regression weights. A second general factor independent of the first, appearing to be a difference between clerical and mechanical abilities, added only about another three percent to the 45% of criterion variance predicted by the first factor. The first general factor was 15 times as effective as the second. However, the two factor scores together accounted for more than 96% of the criterion variance that could be predicted by weighing the tests specifically for each training program.

In the second data set, I sought out data on actual on-the-job performance. The best set of data that I was able to find meeting my rigorous conditions of two independent samples, each composed of at least 50 cases and each validated against some criterion of actual on-the-job performance was in the Technical Manual of the U.S. Employment Service General Aptitude Test Battery. Though the U.S.E.S. has reported studies of over 400 different jobs, there were only 29 of these that met the two criteria I have just specified.

The results for these 28 are summarized in the right-hand column of Table 3. In these data, based as they were on relatively small samples, there was a very marked shrinkage in validity from the original to the cross-validated sample. The average validity in the cross-validation groups was less than half that in the original groups on which the weights were determined. The general g factor score was actually 20% better than the regression-weighted composite. This result was limited to the cognitive tests, but comparable results were obtained for the three motor tests in the GATB. With samples of this size, typical in the industrial psychology literature, one is apparently better off simply to use a measure of general ability and forget about carrying out a special validation study for each job.

This last statement is rank heresy, flying as is it does in the face of the doctrine that tests need to be validated specifically for each job, and that there is a distinctive "best" combination of tests for that job. But I am not alone in that heresy. Schmidt and Hunter, and their associates, have re-examined the validity data for large volumes of civil service tests, for the GATB data, and for results from the AFSAT (the military classification battery). They have undertaken to account for the variation that could be expected to occur from one group to another just by chance in sample sizes encountered

in much personnel research, where 65 cases is fairly typical. Further, they have tried to make some reasonable allowance for differences in the way and extent to which the range of ability has been curtailed in different samples and of the variation in the nature and the reliability of criterion measures. All of these are extraneous factors that could contribute to inconsistent results from one study to another.

Their first meta-analyses were of clerical positions that are found in government service. Here, they concluded that the range of validity values for different tests could be attributed largely, if not completely, to such extraneous factors as have just been mentioned. By implication, if these effects could be eliminated the validity of any given testing procedure would be essentially the same from one job to another. They coined the phrase "validity generalization" to express this conclusion.

Hunter, in particular, has extended the approach to an examination of the GATB data, and to studies of the armed forces classification battery. Within the cognitive domain, he sees most of the potential for prediction being encompassed in one general cognitive ability. This, he believes, is supplemented by a general motor ability, which has its greatest validity in the simpler jobs for which general cognitive ability is least important.

Schmidt and Hunter and their associates are enthusiasts, and may overstate the case for validity generalization. But their analyses provide a healthy corrective to the doctrine of unlimited diversity and specificity. They cause us to recognize that much of the diversity that appears is an illusion, that there is a central core of cognitive functioning that recurs again and again, and that most of the potential for prediction stems from this common core. They lead us to realize that in order to identify with confidence the contribution of factors beyond this common core we must have groups many times larger than those that are likely to be available in civilian personnel research.

You can see from what I have said so far that I am sort of a born-again g-man. But, having brought general ability back to the center of the stage, I do not want to leave the impression that it is the be-all and end-all of academic and vocational prediction. It was only when working with small groups that a uniform measure of general ability outstripped a battery tailored for the specific job, and research with the large groups that were available in military settings indicated the fruitfulness of tailoring a test battery for a specific job -- such as airline pilot. But groups are needed for validation studies that are of a size rarely available in civilian personnel research.

But what is this g, this general ability, that looms so large in human affairs. Attempts to pin it down have ended in more confusion than enlightenment. The often repeated statement that "intelligence is what intelligence tests measure" is an indicator of our frustration in trying to get at its fundamental nature. Up until now we have known it largely through its manifestations in human behavior. We have sought to understand it by studying its correlates in society and in individual lives.

In the last 20 years or so, associated with the flowering of cognitive psychology, there has been a move to examine in detail the processes of thinking and problem solving, and of individual differences in these processes. This approach uses an information processing computer system. This has led some investigators to focus on the limited capacity of working memory to encompass more than a very few thoughts at any given moment, and measures of differences in memory span have shown themselves to be moderately loaded with g. Interest also has focussed on speed of information processing. By a series of ingenious experiments Sternberg dissected the process of responding to analogy items of the type "Cat is to kitten as dog is to ----" into the time spent on assimilating each element of the relationship. He found that the more capable individuals tended to spend a greater fraction of their time digesting the relationship between the first two terms while the less capable tended to jump quickly to the third term, this is, to jump to conclusions, perhaps prematurely.

Jensen and his students at the University of California have repeatedly found a relationship between speed of responding to quite simple stimuli -- such as a choice response to one of a set of lights -- and conventional test measures of g. These studies suggest that an individual's level of g is a reflection of some simple aspect of the efficiency of neural functioning.

There have been further efforts to explicate individual differences in g in terms of individual differences in the physiological functioning of the central nervous system. With the development of more sophisticated and sensitive devices for picking up and recording electro-chemical responses of the brain it has become possible to relate individual differences in events at this level to differences in performance on conventional intelligence tests. Research is still spotty, but some reported relationships have been quite dramatic. These results are in need of replication and confirmation. However, we begin to have the possibility of generating a neuro-physiological theory of the underpinnings of intelligent behavior, one that is biological rather than sociological.

These efforts to dig back to the simplest biological bases of g may eventually lead to understandings that will be a useful guide to social and national policy, but such understanding is still in

the realm of the possible rather than the actual. For the present, we must be content to recognize the reality of g, and its importance as a determiner of the individual's role and effectiveness in our world of work and life.

Table 1. Illustrative Factor Analysis

Analysis A Primary Ability Factors

	Factor I	Factor II	Factor III	Specific	Error
Vocabulary	76	33	33	20	41
Sentence Comp.	79	30	30	17	41
Verbal Classif.	78	31	31	20	40
Verbal Analogies	62	42	48	29	35
Quantitative Rel.	40	65	43	30	38
Number Series	31	67	45	27	42
Equation Bldg.	28	66	22	49	45
Figure Classif.	29	40	67	38	40
Figure Analogies	32	38	67	27	38
Figure Synthesis	27	37	65	44	42

Analysis B Hierarchy of Abilities

	G	V	Q	NV	Specific	Error
Vocabulary	70	51	-02	-02	29	41
Sentence Complet.	74	50	02	-02	19	41
Verbal Classif.	73	50	00	00	24	40
Verbal Analogies	85	30	04	03	25	35
Quantitative Rel.	86	02	21	-02	27	38
Number Series	84	-05	21	04	26	42
Equation Bldg.	73	-03	18	-03	48	45
Figure Classif.	72	00	02	29	40	40
Figure Analogies	83	-02	07	28	29	38
Figure Synthesis	72	-04	00	32	45	42

Table 2. Characteristics of Three Tests of CogAT

Section A - Reliability and Intercorrelations

	Reliability	Quant	Non-Verbal
Verbal Battery	.917	.728	.676
Quantitative Battery	.846		.739
Nonverbal Battery	.857		

Section B - Pooling for Estimate of "g"
Weights to maximize correlation with "g"

Verbal	.82
Quantitative	.89
Nonverbal	.83

Correlation of composite with "g"	.944
Retest reliability for composite	.941

Section C - Components of Variance for Difference Scores

	V vs Q	V vs NV	Q vs NV
Common or "g" factor	72.8%	67.6%	73.9%
Differential factor	15.4	21.1	16.2
Measurement error	11.8	11.3	14.8
Reliability of difference measure	.564	.651	.523

Table 3 - Prediction from Regression Weighted Composite and from Uniform Estimate of "g"

			Army Battery vs. Tech school	G.A.T.B. vs job performance
1.	Weighted composite Own Group	R	.748	.458
		R^2	.560	.210
2.	Weighted Composite Crossed Group	R	.701	.318
		R^2	.492	.101
3.	Uniform "g" Composite	R	.668	.348
		R^2	.446	.121
4.	(3) / (2)		91%	120%

* * * * * * * * * * *

TESTING SOCIAL WORKERS: A CRITERION-RELATED VALIDATION STUDY USING A MULTIPLE TEST BATTERY AND NINETEEN JOB PERFORMANCE DIMENSIONS

Mitchell Drabik
Department of Administrative Services,
State of Connecticut

The testing of social workers has typically been one of the most difficult areas for personnel assessment professionals. Part of the difficulty exists because social work is both a science and an art. Another reason is that social work as a practice depends in large part on the current situation and the social worker's general theoretical orientation.

How then does one adequately test entry level social workers' skills or any social workers' skills. Typically most states and/or municipalities (inclusive of Connecticut) have used some kind of a written multiple choice format designed to test for certain basic knowledge required at entry level such as knowledge of normal human behavior and development, knowledge of sociology and psychology, etc. These multiple choice questions usually require the candidate to choose some course of action from amongst four or five options based upon a capsule-size situation and without much background information. Other critical skills such as written communication, problem-solving ability, assessment and listening skills and interest in the profession have been either ignored or left to the interview situation.

The development and validation of a new entry level examination for social workers was done with the objective of testing for a broader range of social worker skills. It was also done to accomplish the following objectives: (1) to develop as a job-related an exam as possible; (2) to develop a face-valid applicable to social workers in three different agencies (Departments of Children & Youth Services, Mental Health & Human Resources); (4) to develop an exam that is as culturally fair as possible; (5) to provide employing agencies more detailed information about candidates performance for the purpose of making better selection decisions; and, (6) to develop an exam where job content would be changed periodically rather than changing or developing new items.

Job Analysis

The development of a new examination began with a very lengthy and detailed job analysis. The job analysis phase included a series of job audits with incumbents, supervisors and directors of social working each of three employing agencies (total of 9 audits; 3 per agency). A job analysis questionnaire was develo-

ped and issued to a total of 119 social workers encompassing 5 different job levels (social worker trainee, caseworker, social worker, psychiatric social worker assistant, psychiatric social worker). The questionnaire had a task section and a knowledge, skills, abilities, and personality (KSAP) section. Return rate on the questionnaire was 90%.

The job analysis phase yielded 7 distinct job factors. These factors were: Assessment/Problem-Solving Ability, Knowledge of Individual & Group Management Skills, Perceptual Skills, Intervention/Interpersonal Skills, Work Management factors resulted from a statistical analysis of the questionnaire data which listed all critical job tasks and knowledge, skills, abilities and personality characteristics.

EXAMINATION DEVELOPMENT

The exam development phase basically started from scratch. It began with the objective of finding different approaches to testing each one of the seven factor areas - an idealistic goal for sure. The basic strategy that evolved is listed below.

JOB FACTOR	EXAM MODE
Assessment Skills/Problem-Solving Ability	Case scenarios
Knowledge of Individual & Group Behaviors	Case scenarios
Communication Skills	Note-taking Exercise/Essay
Intervention/Interpersonal Skills	Essay Questions
Work Management	Case Scenarios
Perceptual Skills	Group Embedded Figures Test
Personal Orientation to Work	Vocation Interest Inventory

The need to develop as job-related an examination for a Social Worker Trainee was the motivating force to find a different approach to testing social workers. What we came up with a case scenario approach. For this exam we developed three quasi-real cases, one for each one of the three employing agencies. (Depart of Children & Youth Services, Mental Health and Human Resources). The case scenario approach takes the candidates through 3 main phases of the client-social worker interactive process. These are the assessment phase, the treatment or service planning phase and the discharge or termination phase.
Candidates are provided with information as they are typically written in client service records. Blocks of information are

presented to coincide with each of the three main client-social worker interactive phases. In-take information concerning client and family background is presented for the assessment phase. Behavior observations and progress notes are presented during the service planning or treatment phase. Additional progress information and a Community Resource Directory are given at the termination or discharge phase.

The Community Resource Directory is a directory of 16 social work resources available to clients and families throughout the State of Connecticut. Each resource describes the types of services provided, the fees, the target group and the geographic area served. This type of directory had been used successfully with social service workers in the City of Kansas City, Missouri, (Jacobson, 1983). This then was the basic model that was used for exam development.

The development of each case scenario and corresponding questions took a number of sessions. It should be mentioned that both multiple choice and sentence completion items were developed for each case scenario. The use of sentence completion items had not been tried with any other Connecticut state exam.

The case scenarios were developed with the intention of testing candidates' assessment skills/problem-solving ability as it relates to social work situations, knowledge of individual & group behaviors (carry over from previous test), work management skills and some intervention/interpersonal skills. There were 24 questions developed (21 multiple choice, 3 sentence completion) for the Department of Mental Health case scenario, 26 questions (23 multiple choice, 3 sentence completion) for the Department of Children and Youth Services case scenario and 20 (17 multiple choice, 3 sentence completion) for the Department of Human Resources case scenario.

Communication skills (more specifically listening skills) were tested using a note-taking skills exercise. This exercise involved the playing of an eight minute long cassette tape immediately after the case scenarios. The tape consisted of three situations involving the clients from each of the three case scenarios. Test validation participants were asked to take notes during the playing of the tape. They were told that they would be given multiple choice questions based upon their notes later on in the test. The essay part of the examination followed the playing of the tape so that the exercise would not be a memory test. Written communication skills were tested using two essay questions. (One question asked participants to explain why they chose social work as a profession, the other asked them to explain how they would handle a particular social work situation involving a client and their family). A 5 point rating scale was developed to assess their grammar, paragraph and concept formation, etc.

The situation assess question was designed to assess participants intervention/interpersonal skills. Similarly, a 5 point scale was developed by subject matter experts to differentiate amongst participants' ability to handle this situation.

The vocational interest inventory (Bruce Davey, 1983) was used with the intention of identifying some common work interests and preferences of social workers. The VIQ had been found useful with other candidate groups such as State Police Trooper Trainees and Correction Officers. Participants were asked to respond to an 60-item inventory of activities using a Likert type scale ranging from "like extremely well" to "dislike".

The remaining test factor (perceptual skills, a nebulous one at best) was assessed using Witkin's Group Embedded Figures Test. For those not familiar with this test, it is a test of perceptual skills. Perceptual skills are assessed by having the subject locate a previously seem simple figure. Research has indicated that this test goes beyond assessing perceptual skills into other areas of psychological activity such as intellectual functioning, sense or self and body concept.

Criterion Measure

The other major component of this validation project was the criterion measure. The development of the criterion measure was undertaken immediately after the development of the test factors and the linking of tasks to ksap`s. Another committee of 6 social work representatives (i.e. social work supervisors and directors of social work) was formed to identify key performance dimensions in the social work profession and that were directly tied to the test factor.

There were a total of nineteen dimensions that resulted. These were: Client Assessment, Oral Communication, Written Communication, Stress Tolerance, Learning Ability, Knowledge of Individual & Group Behaviors, Attitude, Dependability, Judgement, Initiative, Problem-Solving Ability, Work Management, Intervention Skills, Agency Centered Requirements, Client Centered Requirements, Perceptual Skills, Interpersonal Ability, Basic Counseling Ability, Overall Performance. A five point rating scale was developed following research conducted using these types of scales with case workers in the City of Kansas City (Dieckhoff, 1984). A grand sum or total of performance dimension was used as a key dimension correlated with performance on the different subtests.

Pre-Test Administration and Data Analysis

The next step in the concurrent validation project was the pre-test administration of the battery with employees from the three agencies. Test administration of the five part examination took

approximately 4 hours. There was a total of 221 social workers, 20 psychiatric social worker assistants and 32 psychiatric social workers.

The relationship between test performance and job performance was assessed using: (a) correlations of performance on different arts of the test and each job performance dimension; (b) correlations between performance on different parts of the test and overall job performance. This was done for the entire validation group, a novice social worker group and an experienced social worker group. An item analysis was used to identify problems with individual multiple choice items and to make some adjustments in items prior to using the final test battery.

Table 1 lists the correlations between the grand sum of performance scores (S20) and each subtest for the entire validation group, the novice group and the experienced group.

The three case scenarios and the note-taking exercise had significant correlations with the grand sum of performance for the entire group. There were some differences between the novice and experienced group on these subtests. Differences in correlations between groups were all non-significant.

The last three variables (VIQFC, BIN, FINAL) are all tied into the selection of the final test battery. The variable VIQFC refers to a forced choice version of the VIQ. The variable labelled BIN is the sum of the 3 case scenarios plus the note-taking exercise plus the force choice version of the vocational interest inventory. The variable labelled FINAL is the sum of the 3 case scenarios plus the note-taking exercise.

The individual correlations between each performance dimensions and different subtests (which are not presented here because of the volume of correlations) did not produce any outstanding findings. Correlations were performed between a forced choice version of the VIQ and overall grand performances for the entire validation group, the novice group and experienced group. A forced choice version of the VIQ was created right after the pre-test administration because of prior success using this type of device with other exams and the anticipated need to reduce exam time with the final test battery. THese correlations which are listed in Table 1 were all significant.

TABLE 1

CORRELATIONS BETWEEN SUBTEST SCORES, BIN AND FINAL WITH GRAND SUM OF PERFORMANCE FOR TOTAL VALIDATION GROUP NOVICE GROUP AND EXPERIENCED GROUP.

Variable	Total Group(N-208)	Novice Group(N=61)	Experienced(N=147)
FINAL	.2663 (p.001)	.365(p.001)	.2363 (p.001)

DCYS	.1719 (p.01)	.2262	.1588
DHR	.1405 (p.05)	.1045	.1556
SENTCOMPL	.0355	.0757	.0170
NOTES	.1437 (p.05)	.0979	.1693 (p.05)
ESSAYI	.0924	-.0326	.1316
ESSAYII	.1170	-.0235	.1587 (p.05)
GEFT	.0454	-.2484 (p.05)	.0924
VIQFC	.3799 (p.001)	.2800 (p.05)	.4201 (p.001)
BIN	.3700 (p.001)	.3240 (p.01)	.2690 (p.001)
FINAL	.2737 (p.oo1)	.3249 (p.01)	.2631 (p.001)

Significance levels are all one-tailed.

Having found significant correlations between the grand sum of performance and for each one of the case scenarios, the note-taking exercise and the forced choice version of the VIQ, we decided to experiment with combinations of the subtest scores and run correlations with the grand sum of performance for the entire validation group, the novice group and the experienced group. The variable BIN listed in Table 1, (combination of the three case scenarios, note-taking exercise and forced choice version of the VIQ) shows fairly high significant validity coefficients across all groups. The variable FINAL (combination of the three case scenarios and the note-taking exercise) also shows significant correlations with grand performance across the groups.

Selection of Final Test Battery and Future Use

The selection of test battery consisted of: (a) the three case scenarios with the sentence completion items having been converted to multiple choice items; (b) the note-taking exercise and the ten multiple choice items; (c) one essay questions (non-scored) but which will be provided to employing agencies as an indication of candidates written communication and intervention skills; and (d) fourteen vocational interests forced choice items (responses not figured total scores but to be tried out on an experimental basis inclusion).

The decision to include the three case scenarios lies solely on the validity data showing significant correlations between these case scenarios and overall job performance. The note-taking exercise has some statistical relationship to the total score and

represents an important task and skill that social workers must have in order to be effective and therefore worth including in the test. The decision to include the essay question in the test but exclude it in the calculating the total score was a compromise of sorts. The employing agencies continue to stress the importance of written communication skill or report writing. However, the lack of a significant correlation of the essay scores with overall job performance was reason enough not to include the essay score in the total score.

The decision to use a limited forced choice version of the vocational interest inventory for experimental purposes was also a compromise situation. It evolved from the particular pairings of vocational interest questionnaire items and the anticipated lack of face validity on the part of social work candidates to the particular pairings. While the correlational data for the forced version is significant, the correlations are based upon a simulation and therefore should be assessed with a predictive group.

In addition to using the forced choice version, we will also be administering the full VIQ to candidates actually employed in one of the agencies. We will then be in a better position to assess the impact of the forced choice version versus the full VIQ and make a final decision of the route to travel.

Finally, the removal of the GEFT (favorite of this author) was clearly based upon the lack of any statistical relationship to performance on any of the job dimensions or overall job performance for the entire validation group as well as the perceived lack of face validity by the validation group. Department of Mental Heath employees were more accepting of the GEFT than any other group.

In conclusion, the final test battery will be implemented next month. This examination is given a continuous weekly basis. We intent to analyze the data after a sufficient sample population is obtained and to determine an appropriate pass point. We will be collecting performance data on those candidates actually employed and use this for a predictive study. The general outlook for this case scenario approach to testing social workers appears to have some merit.

References

Davey, Bruce. Vocational Interest Inventory. Connecticut: Bruce Davey, 1983

Dieckhoff, Foster. Employee Performance Manual. City of Kansas City, Missouri, 1983

Jacobson, Larry. Test for Social Service Worker. City of Kansas City, Missouri, 1983

Pincus, Allen & Minahan, Ann. Social Work Practice: Model & Method. Illinois: FE Peacock Publishers, 1973

Witkin, Herman & Oltman, Philip & Raskin, Evelyn & Karp, Stephen. The Manual for the Group Embedded Figures Test. California: Psychologists Press, 1971

* * * * * * * * * * *

COBB COUNTY'S GRADUATED MERIT SAGA: YEAR #2

Kathleen C. Robinson
Employment Services Manager
Cobb County Personnel Department
Marietta, Georgia

Summary

In 1986, Cobb County, Georgia, implemented a true "pay for performance" merit plan. This paper discusses how the plan was implemented and compares results of the process for 1986 and 1987.

The Performance Appraisal System

The performance appraisal system used in the graduated merit program is a refined version of one developed in 1978 by the Georgia Department of Community Affairs and the Atlanta Regional Commission for local governments in Georgia; Cobb County participated in this statewide project. In 1985, the system was implemented on a trial basis, with no tie to pay. Implementation consisted of writing a Supervisor's Performance Appraisal Manual and manual entitled Scale Definitions of Job Performance Factors, and training all supervisors on use of the new system.

In September, 1986, the Cobb County Board of Commissioners approved the graduated merit program and a common review date plan (all employees are evaluated at the same time each year). The graduated merit program officially became effective in February, 1987, when raises were awarded. Some highlights of the implementation of the merit plan included: training supervisors and department heads, holding employee meetings, implementing a

within-department review and a Personnel Department review of the completed appraisal forms, and implementation of a mid-year appraisal process.

The appraisal system consists of five forms for the following job categories: Professional/Administrative, Clerical/Judicial, Manual/Technical, Public Safety, and Managerial. A 6-point rating scale is used. Each form includes job-related factors which are defined behaviorally by statements describing the 6 rating points. The system emphasizes documentation, which is especially important with a true merit pay program. Types of documentation considered "acceptable by the Personnel Department include: Critical incidents, actual examples of performance, ongoing behaviors, and results obtained. Supervisors are provided with "incident reminder" cards to assist them in their documentation efforts.

Two staff members in Personnel are responsible for reviewing all forms submitted by departments under the Board of Commissioners (referred to hereinafter as the "non-elected officials' departments"). Appraisals submitted by elected officials' departments are not reviewed in Personnel. The elected officials had the option of adopting the graduated merit program or remaining on the 5% merit pay "across the board" program which had previously been in effect for all employees. Only 36% of the elected officials have decided to adopt the graduated merit plan.

The Graduated Merit Program

A merit increase guide is used to determine the percent raise awarded to employees, based on their performance appraisal statistical average. The employee receives a rating 1-6 on the factors relevant to the job as given in the appraisal form covering his job. These ratings are averaged to produce the overall rating, or statistical average.

Procedure

Near the end of the year, the performance of all employees is rated by their immediate supervisors. The completed appraisal forms from non-elected officials' departments are submitted to Personnel for review. "Acceptable" forms are sent on for further processing (input of data into a personal computer, then to payroll for processing of the raise). Forms that are considered to be "unacceptable" are returned to the rater for correction or addition of documentation. After a returned form has been corrected, it is reviewed again in Personnel and then sent on for the remaining processing steps.

In June of each year, the Mid-Year Performance Appraisal Feedback Forms are distributed to all departments. This form provides an opportunity for the supervisor and employee to discuss the

employee's performance during the first half of the rating period and identify areas in need of improvement.

Results

Results were presented in terms of two tables and four figures. A t-test performed on the mean percent of forms for both years was statistically significant at the .001 level, indicating that by the second year of the program, supervisors were doing a much better job of completing the appraisal forms.

Overall averages for elected officials' departments, non-elected officials' departments, and countywide were compared for 1986 and 1987. In 1986, the elected officials had the highest average at 4.8, with 4.6 being the average for non-elected officials' departments; the countywide average was 4.7. In 1987, though, the non-elected officials' departments had the higher average (4.8), which could be attributed to the fact that by 1987, the supervisors in these departments had kept better documentation on the performance of their employees; thus, these supervisors perhaps felt more comfortable giving higher ratings to those whom they felt deserved them. In some elected officials' departments, however, a practice of assigning "blanket ratings" of 4's and 5's resulted in a slightly lower average (4.7) than the year before. The overall county average crept up from 4.7 in 1986 to 4.75 in 1987.

A frequency distribution of the percent of employees at each rating level for both elected and non-elected officials' departments graphically presents the points just made. There are peaks at the 4.0, 5.0 and 6.0 levels for the elected officials' departments, while the results for the non-elected officials' departments show a more normal curve.

The overall average ratings for all non-elected officials' departments were presented for both years. Although there is no significant difference in the means of these two groups of ratings, it was noted that 11 of the 21 departments had a change in the positive direction from 1986 to 1987, while 6 had negative changes and 4 remained the same.

Finally, budget results were discussed. In 1986, the raises awarded resulted in the county "going into the hole" a total of $281,098 (.04% of the total personal services budget). In 1987, raises awarded were under budget by $93,701. The change from 1986 to 1987 was explained in terms of turnover and estimates for 1987 being based on actual statistical averages received by employees in 1986.

Conclusions

Cobb County's graduated merit program may be considered a qualified success. On the positive side, the following were noted:

(1) No court suits have been filed as a result of the program.
(2) Fewer complaints were received from supervisors in 1987 than in 1986 regarding the program.
(3) There is evidence that supervisors are keeping better records for disciplinary and termination decisions, as indicated in Civil Service cases.
(4) The actual awarding of the raises went smoothly both years.

On the negative side, the following were considered:

(1) It is disappointing that only 36% of the elected officials have decided to adopt the graduated merit program.
(2) The issue of controls may need to be addressed in the future.
(3) Some employee dissatisfaction with the system has become apparent, which may indicate the need for more supervisory training.
(4) The appraisal forms may need to be revised again, based on input from supervisors and employees.
(5) The issue of setting performance standards must be addressed.

Overall, however, we are optimistic about the future; we successfully met the challenge of getting the graduated merit program implemented and now look forward to a successful continuation of "Cobb County's Graduated Merit Saga".

* * * * * * * * * * *

ASSESSING PRODUCTIVITY

Marianne Bays
Organizational Consultant
Upper Montclair, New Jersey

During the last decade, the improvement of American organizational productivity has been a "hot" management topic. Many organizations have made it a priority to find ways to improve their productivity and, along with this emphasis, have begun to seek ways to monitor productivity and to assess the impact of the organizational innovations that they introduce.

Those of us with personnel assessment backgrounds have the fundamental knowledge, skills and abilities needed to do productivity measurement research. Like any form of personnel assessment, productivity measurement requires an understanding of jobs, organization and psychometrics. However, productivity measurement presents some new challenges to assessment professionals.

New Challenges

First of all, there is only limited experience to draw upon in many areas of productivity measurement. While methods of work measurement in a production environment are fairly well established, there is far less understanding of most aspects of white collar productivity measurement. Valid and reliable measures of intangible work outputs (e.g., research or professional services) are more difficult to develop than are psychometrically sound measures of tangible work outputs. As the service sector grows, this issue becomes more important.

Secondly, the organizational scope of productivity measurement is often broader than other forms of assessment. Selection and promotion assessment procedures typically affect fewer people at one time than do productivity assessment programs. In addition, productivity measurement is generally more threatening to employees than are other forms of assessment. For these reasons, effective productivity measurement program design and implementation must take organizational culture into account. Without this, productivity measurement program success is likely to be impeded by unanticipated cultural issues that result in organizational resistance.

Third, the explication of an underlying business rationale for the measurement effort is essential to the success of the productivity assessment effort. While few people would argue the business necessity for forms of personnel assessment focused on selection and promotion of capable employees, the business case for productivity measurement has not yet been as fully accepted. Further, in the case of productivity assessment, the business rationale varies greatly from organization to organization. Clearly, measurement of <u>all</u> aspects of work productivity in complex organizations is <u>not</u> feasible or necessary. Methods for determining where productivity measurement has the greatest potential payoff to an organization need to be developed and used.

What is Productivity?

There are no simple answers to this question. Some people view productivity as a function of doing work faster or cheaper (i.e., doing more work while holding costs steady or, alternately, doing the same amount of work while decreasing costs). This view is

compatible with classic work measurement techniques where the assessment focus is on the ratio of work input ($ or time spent) to work output (units produced). Such a view serves us well when two conditions hold: 1) We are dealing with an organization with homogeneous work outputs, and; 2) There is management agreement within the organization that information about the efficiency of production of work outputs will help them to better manage work.

There are many types of organizations, however, in which the workload of concern in managing productivity is heterogeneous and not so readily counted. Organizations with professional staff engaged in a variety of types of projects or working on a variety of different cases fall in this category. Here, we could count the number of projects accomplished or cases processed, but the count would not be an accurate estimate of workload because of the great range of complexity across cases and projects. Further, management in these organizations would probably not agree that measures of efficiency of workload production could provide information of high value to them in managing productivity. Instead, they might view things like customer satisfaction level, employee turnover trends or level indicators. While efficiency is important in most organizations, it is not the primary productivity concern in many organizations.

Table I below presents one scheme for looking at productivity and productivity measurement more globally. The work of the organization is broken into two broad components: 1) activities that directly result in product creation and; 2) more indirect, support processes that the organization must perform in order to accomplish its goals. Each has two aspects of productivity: 1) efficiency and 2) effectiveness. Efficiency consists of doing work cheaper, faster or otherwise "righter". It can typically be measured quantitatively and objectively (e.g., with unit cost ratios). Effectiveness consists of doing quality work or doing the "right" work. This aspect of productivity often requires more subjective measures (e.g., client ratings of the quality of service).

No single measure can provide a full picture of productivity since productivity is multi-faceted in all organizations. Further, organizations will differ in the extent to which any type of measure is meaningful. Management of some organizations have a primary business concern with efficiency of product. Others are more concerned with the effectiveness of their product. Others have a greater need for information with which they can better manage the efficiency or effectiveness of their process. Many have a need for information about more than one aspect of their productivity. It is critical to the success of the measurement program that measures be tailored to the specific productivity information needs of the organization.

Table I

Forms of Productivity Measures

	EFFICIENCY	EFFECTIVENESS
PRODUCT	FOCUS ON COST AND BENEFIT OF PRODUCTS DELIVERED TO CUSTOMERS/CLIENTS	FOCUS ON QUALITY OF PRODUCTS DELIVERED TO CLIENTS AND CUSTOMERS
PROCESS	FOCUS ON COST AND TIMELINESS OF SERVICES AND PLANNING PROCESSES	FOCUS ON QUALITY OF SERVICE AND PROCESS USED TO DELIVER PRODUCTS

Organizational Culture

When designing a productivity measurement program and implementation process, attention to the technical measurement issues alone will be insufficient. The measurement professional might successfully develop measure(s) that reliably and validly capture key aspects and yet still fail in the implementation and institutionalization of the program. To be successful, organizational analysis for measurement program planning should include the following:

1. Identification of stakeholders (i.e., people who are key to your efforts because they supply resources, participation, support, cooperation, etc.) Stakeholders may be management or employees of the organization implementing the measurement program, members of other organizations that use the services or products of the focal organization, customers or clients of the organization, or anyone else with vested interest. These are the people that you need to deal with in order to successfully design and implement a measurement program. They may be your supporters, they may be your critics, they may attempt to quietly block your efforts, or they may be indifferent. Knowing who they are is the first step in being able to manage the organizational culture.

2. Assessment of measurement literacy (i.e., the level of understanding of uses and means of productivity measurement) held by organizational stakeholders. Measurement literacy ranges from low to high within and across organizations. Both low literacy and high literacy can pose problems. Where literacy is low, measurement education will need to be provided, fears will have

to be identified and addressed, and steps will need to be taken to channel organizational energy into activities that positively support the program implementation. Where literacy is high, measurement biases will be the major obstacle to overcome. People with previous experience with productivity measurement often have prejudices for or against particular forms of measurement. Unless these biases are drawn into the open and alleviated, they can lead to an (often subtle) undermining of the success of the measurement program effort. When these biases have been identified, however, they can be addressed through education, and organizational energy can then be redirected in support of the program.

3. Identification of cultural dynamics that can impede successful measurement program implementation. Organizational culture is the system of shared values, rituals, symbols, and language that guide organizational behavior. The type of dominant culture and subcultures in an organization, and any conflicts between these groups, are important to consider in the design and implementation of a productivity measurement program.

There are many different schemes for classifying organizational culture. One with particular value in thinking about productivity assessment is that developed by Deal and Kennedy (1982). This differentiates between four types of cultures, called "Corporate Tribes". Each embodies different values with regard to things like risk, advance planning, independence and speed of action. Each will be described in turn, with particular attention given to its implications for productivity measurement program development.

A. Tough-Guy, Macho Culture: This is an individualistic, high-stakes, quick feedback culture. Police departments, management consultants, venture capitalists, sports and the entertainment industry are all examples of organizations where this type of culture is dominant. Successful "tough guys" like to gamble and can tolerate all-or-nothing risks. They have a need for instant feedback. Cooperation is little valued in these cultures. A productivity measurement program here must recognize that:

-- measures must be oriented to the bottom-line, because nothing else matters
-- the level of measurement should be the individual employee because organizational success depends on the performance, management and reward of individual stars
-- measures must provide fast feedback; moreover, the measurement program itself must quickly demonstrate value

B. Work Hard/Play Hard Culture: Most sales organizations are dominated by this low risk, high feedback type of culture. No one sale can make or break a sales rep. Feedback is inherent in the work itself. The idea of good customer service is also

one that permeates. The party-hard aspect of the culture is the organizational response to employees' need for fun to balance the intensity of the work activity. Contests, meetings, promotions, conventions are all means that such organizations use to try to keep employees happy, motivated and to emphasize the importance of the team. A productivity measurement program here must recognize the following:

-- focus on product and process effectiveness (especially customer satisfaction and product quality) will have more management value than focus on efficiency
-- team measures are most appropriate since no individual really makes a difference

C. Bet Your Company Culture: This is a high-risk, slow feedback environment where employees make big stakes decisions and then wait years before they know if their decisions have paid off. Industries where this kind of culture predominates include capital-goods, mining, oil, investment banks, and the actuarial end of insurance companies. The Army and Navy also fall in this category because they spend billions of dollars preparing for the war they might never have to fight. In this culture, the importance of making the right decisions fosters a sense of deliberateness that results in extremely slow and careful movement. The values of this culture focus on the future and the importance of investing in it. The attitude pervades that good ideas should be given the proper chance for success. Successful people in this culture respect authority and technical competence and work cooperatively with others. Here, productivity measurement program design must recognize that:

-- measurement will be a long-term venture because the time frame for product development is itself so long
-- effectiveness measures are likely to be the most highly valued by management, especially those that focus on improving future business process and product quality
-- most efficiency measures, on the other hand, are likely to be resisted because they run counter to key cultural values of slow and careful movement

D. The Process Culture: This type of organizational culture is characterized by low-risk activity with little or no direct feedback from work efforts. Process cultures put order into work that needs to be predictable. Banks, financial service organizations, insurance companies, large chunks of the government, utilities, and heavily regulated industries like pharmaceutical companies are examples of organizations where this type of culture dominates. The values in this culture center on technical perfection--figuring out the risks and pinning the solutions down to a science. In other words, getting the process and the details right. A productivity measurement program in this type of culture must recognize the following:

-- strong resistance to the measurement program is likely to be

encountered; protectiveness and caution are natural responses to absence of feedback

-- stakeholder involvement in the design and implementation of the measurement program will be especially critical to its success, but the tendency of stakeholders to insist on "perfect" measures may lengthen the time needed to accomplish program design and implementation

-- both efficiency and effectiveness measures of process are likely to be of value to management; measures focusing on product are likely to be perceived as less valuable

No one company fits perfectly into any one of these molds. Different parts of the same organization can exhibit each of the four types of cultures. Still, most organizations will have overall tendencies toward one of the cultures because they are responding to the needs of their marketplace. There are also cases, however, where organizations have two very strong and competing cultures. Productivity measurement programs in such organizations will need to be designed to accommodate both of the cultural types that co-exist. Design and implementation of the program will have to be managed so that the key values of both cultures are accommodated. Otherwise, cultural conflicts will simply be fed and the productivity measurement program will become the scapegoat.

Concluding Thoughts

This paper has argued that in the practice of productivity measurement, the assessment professional faces new challenges--particularly where classic work measurement techniques are inappropriate. It has been proposed that the validity and level of acceptance for a given productivity measure will be highly dependent upon the organization's definition of productivity and view of what information can most help them to manage their workload more productively. These views have been shown to vary widely across (and sometimes also within) organizations.

An understanding of an organization's cultural dynamics will provide important insight into the specific critical success factors for productivity measurement program design and implementation in its environment. Deal and Kennedy's (1982) organizational culture typology has been used to suggest how a cultural analysis can yield valuable information about the meaning of "productivity" within an organization and the type of management information most valued by those within it.

Terrence E. Deal and Allen A. Kennedy, Corporate Cultures. Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1982

* * * * * * * * * * *

THE USE OF VIDEO TECHNOLOGY IN A MULTIPLE CHOICE TEST FOR CORRECTION LIEUTENANTS

Paul Kaiser
Principal Personnel Examiner, N.Y.S.
Department of Civil Service

The Correction Lieutenant Test Plan consisted of the following test instruments:

MEMORY TEST (15 MC Questions) - This portion of the test was designed to evaluate the candidates' knowledge of the rules, regulations and department directives that were determined to be both critical and which the incumbents determined to know cold. That is, the incumbents typically would not have the time to refer to or look up the directives on the job in response to given situations.

The candidates were sent copies of all the directives which fit this paradigm approximately one month before the test and were not permitted to refer to this material during the test itself.

OPEN-BOOK TEST (60 MC Questions) - This portion of the test was designed to evaluate the candidates' knowledge of the rules, regulations and department directives that were determined to be critical but the incumbents typically would be able to refer to on the job in response to given situations. The candidates were provided with copies of all rules, regulations and related material during the test which they could refer to, as needed, when answering the questions.

VIDEO TEST - (15 MC Questions) - This portion of the test was designed to present the candidates with non-written test material which would evaluate skills and abilities that could not be measured in other components of the examination. The hypothesis was that a non-verbal test situation presentation would have less adverse impact on protected class candidates. The candidates were presented six video scenes and were referred to specific questions in a test booklet which they had to answer based upon their understanding of the video scenes presented.

INCIDENT SIMULATION TEST (4 Problems) - This portion of the test was designed to evaluate the candidates' higher level decision-making and analytical skills and abilities that could not be otherwise evaluated in other components of the test. Problem one was designed to present the candidates with an emergency problem; problem two presents a stabbing investigation situation to consider; problem three was a supervisory problem; and problem four was a series of "day-in-the-life" situations that the candidates had to deal with.

Video Script: Scene #1: "The Senator's Wife"

SITUATION: In this scene, you will be shown an interaction between a Correction Sergeant, a Watch Commander, and a Senator's wife.

LOCATION: Watch Commander's Office; facility entrance gate.

CHARACTERS: Watch Commander, Correction Sergeant, Senator's Wife.

(Location: Facility Entrance Gate)

SEN'S WIFE TO CORR SGT: "I don't understand the problem here. I was invited here by the Superintendent to give a presentation to the inmates of this facility. Now, you're telling me that you expect me to undergo the indignity of a metal detector search and I'm supposed to dump my purse onto the table for you to inspect the contents? I have personal property in this purse; I am not going to dump my purse. Who do you think you're speaking to? Do you seriously think I'm smuggling contraband into this facility?"

CORR SGT TO SEN'S WIFE: "I regret having to ask you to do this; however, no visitors are allowed to enter the facility without going through the search that we're asking of you. We're not asking you to go through an extensive search of your personal clothing. All we're asking is that you allow us to hand-scan you with the metal detector and then allow us to examine the contents of your purse. We're not asking you to do anything that we wouldn't ask of other visitors to the facility. The requirements are very clear on this point. We are only asking you to do what's required by the regulations."

(Scene shifts to Watch Commander's Office.)

CO TO WATCH COM: "Lieutenant, the Gate Sergeant called to say that he's having a problem processing one of the visitors who happens to be a Senator's wife. We need you to come down to the gate."

(Scene shifts back to gate area)

SEN'S WIFE TO SGT: "I resent your attitude! I am not just anyone! I won't dump my purse, and I suggest you get the Superintendent down here and tell him I'm here to make my presentation."

(Watch Commander enters gate area.)

WATCH COMM TO SEN WIFE: "Excuse me, my name is Lieutenant Oliver and I'm in charge of this facility at this time. The Superintendent is off; it's after duty hours. What seems to be the problem?"

CORR SGT TO WATCH COMM: "Lieutenant, this is Mrs. Ryan, Senator Ryan's wife. The Superintendent has invited her to make a presentation before the inmates, but she won't submit to the required searches."

SEN'S WIFE WATCH COMM: "The Sergeant isn't listening to me. I was invited by the Superintendent to give this presentation. I am not smuggling contraband. I do not dump my purse. I'm not just anyone!"

WATCH COMM: "Ma'am, the Sergeant was correct in not allowing you to enter this facility without undergoing a routine search. This is nothing personal, and we're not in any way implying that you've got anything to hide. But I hope you understand that what we're trying to do is to simply follow the regulations that have been established by our department to maintain the integrity and the security of this facility."

SEN'S WIFE TO WATCH COMM: "Well, when I was invited by the Superintendent, I never expected this; and as far as I'm concerned, HE (Senator's wife points at Sergeant) owes me an apology, and only after I get such an apology might I consider your idiotic search!"

CORR SGT TO SEN'S WIFE "I don't owe you anything! I'm here doing a good job, and all you're doing is giving me a hard time!"

Test Items: Scene #1: "The Senator's Wife"

SITUATION: In this scene, you will be shown an interaction between a Correction Sergeant, a Watch Commander, and a Senator's Wife.

1. As Watch Commander, what action would you take at this point?

   *A. Direct the Sergeant to leave the area while you talk to the Senator's wife.

B. Tell the Senator's wife that if she refuses to the search she must leave the facility.
C. Direct the Sergeant to inspect the Senator's wife's belongings.
D. Show the Senator's wife a copy of the directives regarding entrance to the facility.

	Totals (P=.39; Rpbis=.42)				White (P=.43; Rpbis=.42)				Black (P=.27; Rpbis= .32)			
	(A)*	(B)	(C)	(D)	(A)*	(B)	(C)	(D)	(A)*	(B)	(C)	(D)
HI	245	141	2	50	212	99	1	32	27	33	0	13
LOW	98	274	2	64	82	219	1	42	12	43	1	17

2. What action would you take concerning the request by the Senator's wife to call the Superintendent?

A. Call the Superintendent at home.
*B. Tell the Senator's wife that you will pass her request on to the Officer of the Day.
C. Assure her that the problem can be resolved without calling the Superintendent.
D. Tell her that you cannot comply with her request.

	Totals (P=.42; Rpbis=.26)				White (P=.41; Rpbis=.26)				Black (P=.47; Rpbis=.28)			
	(A)	(B)*	(C)	(D)	(A)	(B)*	(C)	(D)	(A)	(B)*	(C)	(D)
HI	11	227	164	36	10	117	130	27	2	28	30	3
LOW	19	143	191	84	15	105	151	72	3	31	30	9

3. What action would you take regarding the Sergeant's handling of the situation?

*A. Verbally counsel him for inappropriate behavior.
B. Verbally commend him for how well he handled a difficult situation.
C. Take no action because no action is necessary.
D. Issue him a formal written counseling memorandum.

	Totals (P=.60; Rpbis=.40)				White (P=.63; Rpbis= 36)				Black (P=.47; Rpbis=.46)			
	(A)*	(B)	(C)	(D)	(A)*	(B)	(C)	(D)	(A)*	(B)	(C)	(D)
HI	331	29	70	8	265	21	42	6	49	11	12	1
LOW	194	99	135	84	167	73	94	10	20	21	31	1

* Correct Answers

Summaries of EMERGENCY SIMULATION By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	9.2584	4.1392	747	
WHITE	9.5693	4.0459	599	
BLACK	7.9483	4.3456	116	.41
HISPANIC	8.1875	4.1226	32	.42

Summaries of INVESTIGATIVE SIMULATION By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	9.5181	1.9759	747	
WHITE	9.7129	1.8254	599	
BLACK	8.6379	2.4473	116	.68
HISPANIC	9.0625	1.8997	32	.35

Summaries of SUPERVISION SIMULATION By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	8.9264	2.0057	747	
WHITE	9.2304	1.8179	599	
BLACK	7.5603	2.3267	116	.85
HISPANIC	8.1875	1.9082	32	.45

Summaries of DAY-IN-THE-LIFE SIMULATION By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	9.6439	2.3487	747	
WHITE	10.0367	2.1253	599	
BLACK	8.0431	2.6748	116	.86
HISPANIC	8.0938	2.0058	32	.75

Summaries of MEMORY SUBTEST By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	13.3369	1.6601	745	
WHITE	13.5059	1.4821	597	
BLACK	12.4914	2.2554	116	.69
HISPANIC	13.2500	1.3440	32	.27

Summaries of OPEN BOOK PART 1 By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	13.7369	1.7104	745	
WHITE	14.0101	1.3548	597	
BLACK	12.5603	2.5066	116	.87
HISPANIC	12.9063	2.0691	32	.55

Summaries of OPEN BOOK PART 2 By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	13.4752	1.7764	745	
WHITE	13.7253	1.5089	597	
BLACK	12.4569	2.4223	116	.72
HISPANIC	12.5000	2.0320	32	.66

Summaries of OPEN BOOK PART 3 By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	13.0564	1.9077	745	
WHITE	13.2764	1.8150	597	
BLACK	12.1897	2.0680	116	.60
HISPANIC	12.0938	1.8554	32	.66

Summaries of OPEN BOOK PART 4 By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	12.8952	1.9509	744	
WHITE	13.1913	1.7141	596	
BLACK	11.6638	2.3809	116	.76
HISPANIC	11.8438	2.3016	32	.64

Summaries of VIDEO SUBTEST By levels of ETHNIC

	Mean	Std Dev	N	Diff*
TOTAL	8.4040	1.7823	745	
WHITE	8.5126	1.7092	597	
BLACK	7.9741	2.0107	116	.33
HISPANIC	7.9375	1.9828	32	.28

*Note: Diff=Difference in Means for Standardized Scores

* * * * * * * * * *

Use of Videotaped Work Sample Material In Interpreter Testing

Michael W. Minter
New York State Office of Court Administration

Background

The need for testing bilingual personnel in the public sector has become increasingly important in the past few years with the great influx of non-English speaking immigrants into many parts of the United States. This is a particularly crucial concern for the judiciary in regard to assuring equal access to the courts for linguistic minorities. The New York State Office of Court Administration (OCA) has addressed this issue primarily through the position of Court Interpreter. In addition to hiring per diem interpreters as needed in more than 50 languages, OCA has 120 full time positions for Spanish Court Interpreters. This presentations focuses on the selection techniques used for the Spanish Court Interpreter title and, in particular, on the use of videotaped material in the administration of the oral portion of the selection exam.

The job of Court Interpreters primarily involves oral courtroom interpretation. They may also do non-courtroom interpreting, such as at hearings, conferences, psychiatric interviews, and defendant/attorney meetings. Court Interpreters do oral translations of written English material, such as charges and waivers of extradition, into Spanish for defendants. Occasionally they make written English translations of material, such as documents from Spanish-speaking countries or of audio tapes from wiretaps of individuals speaking in Spanish.

Testing Strategy

Several issues had to be addressed concerning the development of a testing strategy. First, the exam had to test equally for English and for Spanish. Special attention was paid to which aspects of the language were most important for court interpretation. Accuracy, comprehension and fluency were all important. Vocabulary was a particularly critical issue. An extensive vocabulary was needed - the standard language of educated and professional people; legal, medical, and other specialized terminology; and street and slang terms ("Spanglesh"), including the language of the drug and criminal subculture.

It was clear from the incumbents and the exam committee that testing only for bilingualism was not enough. Oral interpreting abilities must be assessed as well. Therefore, a traditional paper-and-pencil test by itself would not be adequate. There was also practical considerations. It was expected that as many as 2,000 people would have to be tested. The cost and scheduling requirements of an oral exam for such a large group would have been prohibitive. The decision was made, therefore, to give a written test (assessing basic language skills in English and Spanish), which would serve as a screening device. Candidates who were successful on the written test would then be asked to take an oral exam.

Oral Exam

In order to be as job related as possible, the oral exam used a work sample/job simulation approach. Scripts were developed based on actual cases in Civil, Family, and Criminal Courts. In each case there was an English speaking attorney and a Spanish speaking witness. Candidates were required to translate into Spanish everything spoken in English and vice versa.

When an oral exam had been given previously in 1981, the entire process was done "live". Two actors read the scripts to the candidates who did the interpreting in front of a panel of raters. After discussions with several language experts, it was decided to take a new approach for the 1987 exam. A video-tape of actors reading the exam scripts was made. Each candidate was played the tape on a television screen, and simultaneously an audiotape of his/her oral interpretation was made. This tape was evaluated at a later time.

Professional facilities were obtained through New York State Civil Service in Albany for preparing the videotape. The tape was edited so that there were pauses of appropriate length to allow for the candidate's interpretation. In this way the tape never had to be stopped once the exam started. A short practice portion was added to the beginning of the tape. The tape ran for approximately 30 minutes. When the tape was finished, candidates were given two short written passages (one in English and one in Spanish) to review for five minutes and then a sight translation of the passage was included at the end of their audio tapes.

Conclusion

The use of the written screening test and the videotape oral test worked well. For the written exam 325 candidates (16.21%) passed out of 239. The correlation, uncorrected for restriction of range, between the oral and written tests was 0.232 ($p \leq .0003$).

The video oral exam had many benefits. It still maintained the work sample/job stimulation approach, but allowed for more

standardization of input. Previously, when the scripts had been read to each candidate, it was impossible to insure the same rate of speed and same pronunciation throughout several days of testing with different actors. A frequent complaint from candidates in oral exams is that their performance was "unnaturally" worse because of test anxiety. Feedback from the candidates was much more positive in this administration. Several candidates expressed relief over not having to perform in front of a group of people. Although we could not test for it, it was felt that the influence of halo effects and other rater bias was reduced by not having the raters see the candidates. Finally, the cost factor should be noted. Obviously, the costs of oral exams is much greater than paper-and-pencil exams. With the previous exam actors and raters had to be scheduled for each candidate. Late comers and no shows added to the problem. The use of videotapes greatly facilitated scheduling for the raters, the candidates, and OCA. This procedure also provided a complete record of the exam in case of challenges from the candidate about his/her score.

Beginning in the near future, we plan to use this videotape method for screening of per diem Court Interpreters in other languages.

* * * * * * * * * *

EXPLORING A LEGAL DEFINITION OF SUPERVISION AND ITS IMPACT AS A SELECTION CRITERION

Patrick T. Maher, Principal Associate
Personnel & Organization Development Consultants, Inc.
La Palm, California

Abstract

This paper examines the concept of supervision as a job-related element as well as a selection criterion, and provides suggestions on how to address the issue, both legally and psychometrically in validation studies and in assessment procedures.

A work behavior typically critical to first-line supervisory through at least middle management positions -- and often times to varying degrees at the executive level -- is that of "supervision".

In reviewing the legal cases involving this element as well as a number of examinations that have been criticized in a litigious situation, it has become apparent that some assessment specialists are misapplying the elements of supervision.

If a potential for Title VII litigation exists, then the prudent examination developer would most certainly want to anticipate such legal challenges and avoid an invitation to such litigation. Further, it is always prudent to develop a legally defensible examination.

While it may seem somewhat fundamental, it is important to first review the basic concepts -- in particular, legal applications-- required to develop an examination that will be defensible in Title VII litigation.

It is now generally recognized that any assessment procedure that has adverse impact must be validated for job relatedness. A factor that apparently is not as well known to many assessment specialists, however, is that an examination must measure appropriately those attributes critical to the position.

> ...it is reasonable to insist that the test measure important aspects of the job, at least those for which appropriate measurement is feasible...(Guardians, 1980)
>
> To be representative for Title VII purposes, an employment test must neither: (1) focus exclusively on a minor aspect of the position; nor (2) fail to test a significant skill required by the position. (Gillespie, 1985; emphasis added)

This concept is not unrealistic, although it is often neglected. Obviously, if you can only measure job related attributes it follows that it is critical job-related attributes that must be measured.

The courts also require that a content validation study involve certain processes. Among these is the identification of critical work behaviors and the identification of critical knowledges, skills, or abilities (KSAs) linked to one or more specific critical work behaviors (Vulcan Pioneers, 1985; United States Civil Services Commission, 1975; Long, 1981).

The Uniform Guidelines (1978) define a work behavior as

> An activity performed to achieve the objectives of the job. Work behaviors involve observable (physical) components and unobservable (mental) components. A work behavior consists of the performance of one or more tasks. Knowledge, skills, and abilities are not behaviors, although they may be implied in work behaviors.

Supervising subordinate employees, therefore, is a work behavior.

A knowledge, skill, or ability (KSA) must be possessed in order to perform a work behavior at varying levels of competence. What seems to be happening, however, is that this clear distinction is not always made. Thus, supervision is being both identified as a critical work behavior, then reclassified as a KSA, leading to confusion. For example, a work behavior may be defined as "supervise subordinate employees" while "ability to supervise" is identified as the KSA being measured.

In order to avoid this pitfall, the assessment specialist must make a clear distinction. One choice is to only use "supervision" as a work behavior. Then assess the work behavior by developing as selection procedure representative of the behaviors for the job in question, or develop a selection procedure that provides a representative sample of the work product of the job (Uniform Guidelines, 1978).

Or, the assessment specialist must identify and operationally define the critical KSAs necessary to perform the various supervisory work behaviors. These critical KSAs must then be evaluated in the assessment procedure (Uniform Guidelines, 1978).

It is also important to realize that "supervision" does not consist of a few elements. Depending upon the specific job, supervision can entail a number of different work behaviors. For example, if a supervisor must prepare performance evaluations on a subordinate, this task can generally be identified as a separate work behavior or work behavior cluster involved in supervisory functions. Likewise, if the supervisor must conduct investigations into allegations of improper work performance, whether such investigations are formal or informal, then such investigations can usually be considered another distinct supervisorial work behavior. Other activities, such as scheduling and training personnel, inspecting or reviewing work, and making work assignments all might fall under the broad umbrella of supervision. Again, successful performance of each of these distinct work behaviors will require a number of KSAs, although many, if not all, of these KSAs may be identical from work behavior to work behavior.

As an example, we can look at the following description of work behavior:

> Investigates allegations of misconduct, inattention to duties, or poor service, determines the validity of complaints, and, where necessary, prepares letters of reply, memoranda, or other appropriate documents.

The following KSAs are some that might be identified as critical to successful performance of that work behavior:

Knowledge of the rules, regulations, policies, and procedures of the department

Knowledge of court decisions and statutes affecting disciplinary actions

Knowledge of contemporary management and supervisory procedures and principles

Ability to orally express ideas, tasks, directives, conditions, needs and information, concisely, accurately, clearly, and persuasively

Ability to identify problems, evaluate courses of action, develop alternative courses of action, and reach logical decisions based on the information at hand

Ability to perceive and react to the needs of others

Ability to clearly and effectively express ideas in writing

By measuring these KSAs in a variety of assessment procedures, we can then determine the extent to which a candidate possesses them and can likely predict his performance of the work behavior on the job.

Problems develop only when supervision is viewed in and of itself as the work behavior being performed and further is translated into a KSA. When this happens, there is likely to be an inferential leap to measuring supervision as a KSA, which is not only inappropriate, but will likely lead to an inappropriate procedure. This exact situation was ruled improper in Vulcan Pioneers (1985). The assessment specialists attempted to measure supervision as a KSA strictly through a paper-and-pencil test. The court, not surprisingly, found that supervision involved more than correctly answering multiple choice questions and that other assessment procedures were necessary.

Since supervision invariably involves oral communication skill or abilities, interpersonal relations, and perhaps other elements, it is obvious that it cannot be measured simply through a job-knowledge test. There is no doubt that knowledge of certain supervisory principles, theories, or practices relevant to the ability to perform supervisory tasks properly can be adequately measured on a job-knowledge paper-and-pencil test, but such knowledge is only one aspect of successful performance as a supervisor.

In summary, measurement of supervisory and, by the same token, management, work behaviors cannot be artificially narrowed if one is to develop a legally-defensible and, indeed, professional assessment procedure. An attempt to measure the multitude of KSAs necessary to the performance of critical supervisory work behaviors through a simple paper-and-pencil test of knowledge has not been accepted by the courts and is not likely to be.

The key to a content-valid, defensible assessment procedure is to thoroughly analyze and identify the critical work behaviors in the supervisorial function and then identify critical KSAs. The final step is the development of an assessment procedure that properly and adequately measures those critical KSAs.

Obviously, the nature of supervision is complex enough that a paper-and-pencil test will never suffice as the sole assessment procedure for either work behaviors or their underlying KSAs.

References

Long, J.E., The rater's guide to KSA-based job analysis. Seattle, WA: U.S. Office of Personnel Management, 1981.

United States Civil Service Commission, Job analysis for improved job-related selection. Washington, DC: U.S. Government Printing Office, 1975

Gillespie v State of Wisconsin, et al (1985) 771 F3d 1035

Guardians Ass'n of New York City v Civil Service, (1980) 630 f.2d 78

Vulcan Pioneers v New Jersey Department of Civil Service, (D.N.J. 1985) 625F.Supp 527

Uniform Guidelines on Employee Selection Procedures, (1978) 43 CFR 166

* * * * * * * * * *

WHOEVER IS REASONABLY PROFICIENT IN THE WORK PLACE, PLEASE RAISE YOU HAND!

Eileen A. Groves, former Assistant City Attorney
Columbus, Ohio
Associate Corporate Labor Council
Borden, Inc. Columbus, Ohio

The Uniform Guidelines provide that:

> Where a cut-score is used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency with in the work force.
> 29 C.F.R. Sec. 1607.5 (H)

Industrialist psychologists recognize the organizational performances range along a continuum between high and low reflecting proficiencies of employees. When an employer is seeking to hire or to promote, it is his aim to improve or raise the quality of his employees or his supervisors. Organizational performances can be improved or raised by improving the selection and training of employees and supervisors. Courts, especially federal courts under Title VII and other anti-discrimination statutes, do recognize generally that organizations seek to function properly and efficiently. But it must be recognized that, under discrimination statutes, selection devices or promotional devices which have an adverse impact upon protected groups become suspect. If a selection device does have an impact, the device itself and the cut-point become suspect.

The key question to the establishment of an "acceptable" cut-point is:

> What is reasonable and consistent with normal expectations of acceptable proficiency?

In Columbus, we have gone through, in the past ten years, a series of testing cases involving our public safety officers. In Brant v. City of Columbus,[1] there was a challenge to the police selection testing devices which include a physical agility test. The Court in 1979 indicated that any test with a cut-score that would have eliminated 30% of the incumbents is error. Cut-scores should not eliminate incumbents unless there is a clear demonstration that they're not performing satisfactory. In 1986-1987, in Brunet v. City of Columbus,[2] a case which I have been involved with since early in 1985, the court mandated in its own interim scoring scheme that a cut-score should be at one standard deviation below the incumbent mean on the physical test overall or the point at which 16% of the incumbents would have failed.

During the time in which the City was presenting evidence to the trial court in the firefighter case, the City was also in discussions with the federal court as to the police promotional examination. This court, a different judge, simply suggested that the City apply what had been the traditional Civil Service cut-score. The District Court in the police promotional indicated no reasoning for its decision to accept 70% as a pass/fail point aside from this was the historical point. This inconsistency is illustrative of the history of cut-scoring within the case law.

Since the passage of Title VII, there have been many testing cases which have risen to challenge examination validity and job relatedness. Courts, on occasion, when examining validity and job relatedness have looked to scoring and ranking because of the adverse impact that a scoring scheme has upon minority groups. In 1973, in Bridgeport Guardians v. Bridgeport Civil Service Commission, 482 F.2d 1333 (2nd Cir. 1973), the court criticized as archaic a 75% pass/fail rule. The court specifically found that this "arbitrary determination was indicative of an archaic testing system, particularly where there was no evidence of weighing of questions based on actual job requirements." Subsequently, in 1979, in the case of Association Against Discrimination v. City of Bridgeport, 594 F.2d 306 (2d Cir. 1979), the District Court found that the ultimate effect of the examination turned on the score used to differentiate between passing and failing. Under Bridgeport City Charter, all candidates had to answer correctly as least 75% of all questions on the Civil Service examination. The District Court characterized this application as having no relationship to job proficiency, particularly when a consulting firm who prepared the examination did not recommend a passing score. The defendants in a remedy proposal urged the District Court to lower the passing score thus eliminating most of the disparate effects of the examination. The Court of Appeals found the City's arguments persuasive and reversed the case and remanded it to the District Court for consideration regarding the passing score proposal of the defendants.

In a subsequent appeal following remand, the Second Circuit noted in Association Against Discrimination v. City of Bridgeport, 647 F.2d 256 (2nd Cir. 1981), that the new score would have an adverse impact upon minorities though not as substantial. The Court ordered affirmative relief.

In Guardians Assoc. of the New York City Police Dept. v. Civil Commission of the City of New York, 630 F.2d 79 (2nd Cir. 1980), the City of New York used the results of the examination to compile a rank order list of all applicants and then selected a passing score which sufficiently generated the required number of potential recruits. The Court of Appeals held that neither the rank ordering or the passing score selection conformed to the

minimal professional standards. The Court held that the relationship between higher scores and better job performances might permissibly be inferred, but, where the test scores reveal a disparate impact and the disparity is greater at high passing scores that at low passing scores, the appropriateness of the inference of higher scores with better job performances must be closely scrutinized.

After its discussion of rank ordering, the Second Circuit embarked upon a discussion of cut-off scoring. The Circuit indicated that there should be some independent basis for choosing the cut-off point. A criterion-related study would not be necessarily required if the employer established a valid cut-score by using "professional estimates to locate the logical "break-point" in the distribution of scores." The Court held that if it had been demonstrated that the examination measured abilities with sufficient differentiating power to justify rank-ordering, it would be valid to set the cut-score at the point where rank ordering filled need.

It must be noted that both in the Bridgeport cases and the New York Guardians cases, the Circuit Court repeatedly went back and looked at the results and indicated that employers could look at test results. It would also appear that the Second Circuit was amiable to lowering of pass scoring if it would diminish or eliminate adverse impact. Very recently, however, the Ninth Circuit in San Francisco Police Officers Assoc. v. The City and County of San Francisco, 812 F.2d 1125 (9th Cir. 1987), held that the City Civil Service Commission action of reweighing examination components on a promotional examination impermissibility trampled the interest of non-minority police officers where the Commission knew the candidates' race and gender and how the candidates performed in individual test components when they made the decision to alter the examination pattern. Use of an alternative selection procedure was unlawful because it permitted the Civil Service Commission to manipulate the results to produce the desired racial and gender percentages.

In Burney v. City of Pawtucket, 559 F. Supp. 1089 (D.R.I. 1983), the Court in its decision found that the cut-scores were arbitrarily extracted by the City's decision to eliminate at the 15 percentile of men. The Court found that this flew "in the teeth of the Guidelines, which require that cut-off scores be 'set so as to be reasonable and consistent with normal expectations of acceptable proficiency.'"

In Thomas v. The City of Evanston, 610 F. Supp. 422 (N.D. Ill. 1985), the District Court found no empirical evidence to support the assumption that 16% of the incumbents were physically incapable of performing the job. The Court in its final decision concluded that there must be some evidence to support a principle decision that a cut-off figure really predicts job performance.

There should generally be some independent basis for choosing the cut-off. Does this mean concurrent criterion-related validity studies?

More recently, the question of cut-off scores and ranking in fire has taken various tracks with inconsistent results. The most famous, or infamous, fire testing case is Berkman v. City of New York. Initially filed in 1979, the plaintiff alleged gender discrimination challenging the physical entrance test for the New York Fire Department. In 1982, the District Court invalidated the physical portions of the New York Fire Department examination and ordered several forms of relief, including the preparation of a valid selection procedures, affirmative hires and the conducting of a validity hearing. There was a criterion-related validity study done comparing test scores with the job performances. The experts testified that the analysis showed a high degree of correlation for both male and females between physical test and the job. The plaintiffs, however, challenged the proposed banding of scores, and wanted broader bands of candidates with random selection within the bands.

The Second Circuit in February 1987 affirmed the District Court's finding that the physical examination was job related and content valid. The Appeals Court rejected the three-band system as neither enhancing the validity of the physical test nor reducing the adverse affect on women. It is noteworthy that the only evidence as to the scoring was the indication that the criterion related validity study compared the test scores of incumbents with their job performance and that there was a high degree of correlation between the physical scores and job performances. There was no discussion, however, as to the cut-point or the differentiation on the scoring bands. It can only be presumed that the studies supported the bands and cut-point. More recently, in May of 1988, in the case of Barbara Zamlen v. City of Cleveland, the United States District Court found that the Cleveland Fire entrance examination, particularly the physical entrance examination, was job related and content valid. The District Court in Zamlen simply indicated that the City's Charter provisions and Civil Service regulations provide for examination, testing and hiring by rank order of City employees. The Court held there was nothing improper with this decision so long as the procedures used were not discriminatory against minorities and women. It held that the City could make a policy decision to hire the best qualified and provide for rank ordering so long as it did not discriminate.

The Court found that the concurrent validity studies revealed that firefighters who scored highest on the examination did better on the job. This District Court did not discuss this implication but simply accepted the 70% cut point as provided for within the City Charter.

As I indicted in the opening, there is no clear direction from courts to the conflict in test creation and evaluation. This is most clearly evident, I believe, in Brunet v. City of Columbus, a case that I am most familiar with. In Brunet, a group of four female applicants challenged both the written and physical entrance examinations to the Fire Department. Subsequent to the initial trial court decision, the Court found that the 1986 examination was content valid and job related if the defendants omitted the hose hoist event. The Court felt this event had appreciable adverse impact upon women. The Court, before issuing its findings, had specifically requested the defendants to recalculate the results of the tests with the omission of the hose hoist and report to the Court.

The defendants had pretested the 1986 examination using a group of 145 firefighters ranging in the age from 22 through 57 and calculated their means and standard deviations. The defendants also conducted an analysis of the variances on the scores and determined that .89 of the variance was attributable to age. When you exam the defendants' scoring proposal, approximately 65% of the current firefighter sample would have been able to perform all the test events successfully. Recognizing that the applicants were in the 20 to 29 year-old bracket, a comparable ample of the subgroup of the incumbents indicated that 94% of the incumbents would have passed all the test events.

As I indicated in the beginning, there is a conflict between concerns of industrialist psychologists, employers, and the courts. In the case of the City of Columbus in Brunet, content validity studies, criterion-related validity studies and other content validity professional studies to justify both its physical and medical screening, costed nearly $300,000. Quite frankly, we are currently also facing a request by the plaintiffs for nearly a half-million dollars in attorneys fees because they were successful in getting a court order for two women to be hired. Can small municipalities or small employers afford to spend several hundred thousand dollars creating and proving their employment devices?

How do we address these concerns and conflicts? I can offer you no answers. There does have to be cooperation between industrialist psychologists, employers and the courts. There must be realism and practicality. But the topic of this presentation is: Who is reasonably proficient in the work force? - I can offer you no answers, can you give me any? No court has yet indicated how it is defined.

---

[1]Unpublished.

[2]642 F. Supp. 1214 (S.D. Ohio 1986), dism'd as moot, (6th Cir. 1987), cert denied, U.S. ___(1988).

* * * * * * * * * *

SCREENING DIRECT CARE WORKERS FOR CHILD ABUSE POTENTIAL

Martin W. Anderson
State of Connecticut*
Department of Administrative Service

*This study was conducted while the author was Director of Personnel Assessment for the State of Oklahoma Office of Personnel Management. Significant contributions in data collection were made by Leonard Anderson, Sara Bohanon, Joe Davenport, Robb Hayes, Vivian Pegues, and M.M. Sundram.

Background

There has been recent attention and concern regarding the quality of care offered to institutionalized children and adults by state facilities. Of greatest concern is the abuse and neglect of these state clients. This is a topic which has not escaped the notice of the professional literature (See Volume 42, Number 11, of the American Psychologist, 1987)

The state of Oklahoma has come head to head with litigants challenging the quality of care for children in her custody. This has been most pronounced in recent court action seeking the removal of mentally and multiply handicapped from an institution setting and placing them in group homes. Plaintiffs claimed the institution had unsanitary living conditions, a lack of proper habilitation programming, segregation form the larger society and maltreated the children ("The Hisson Struggle", Tulsa World, May 16, 1988, p.11). The court found that the facility must be closed and all clients placed in group homes within four years.

Litigants and federal laws have placed unusual burdens upon resident care facilities. For example, an ombudsman must be available to all clients. The purpose of the ombudsman is to have someone in the facility to whom clients can report any incident which they consider to be maltreated. The human services agency keeps records of these reports. From a sample of reports collected within a twelve month period of time, 83.6% were labelled as abuse claims, 9.4% were labelled neglect, and 7.1% were labelled as mistreatment. The human services agency believed more needed to be done to screen employees who would work with their clients.

The principle employees cited as the most troublesome in abuse/-maltreatment/neglect cases were Resident Life Staff Aides (RLSAs). These are employees who have the most direct contact

with facility clients. The RLSAs are responsible for providing direct care ranging from toileting to habilitation programs to transporting some higher functioning clients to paying jobs. Resident Life Staff Aides are designated as a noncompetitive classification. That is, no formal tests or assessment devices are taken by applicants before they are hired.

The human services agency approached members of the Oklahoma Office of Personnel Management regarding ways in which to better screen RLSA applicants. Of greatest concern to OPM management was the fact that the agency wished to administer RLSA applicants the MMPI in an attempt to screen for or diagnose behavior tendency. This was seen as inappropriate and unsupported from both a clinical and personnel assessment standpoint. OPM suggested something more focused and job related be considered after a job analysis.

Job Analysis

A job analysis was conducted on RLSA incumbents to investigate what components could make up a systematic selection scheme to screen applicants. The job study began with two major efforts. First, a comprehensive inquiry was initiated wherein significant management and supervisory personnel were interviewed regarding the administrative problems caused by RLSAs. Second, a comprehensive job analysis was conducted on RLSAs.

A number of important findings came from the inquiry into the administrative problems caused by the RLSAs. The major concern was keeping people ill suited for working with such limited and defenseless clients from being employed. Another important concern was the literacy skills of RLSAs. With the litigation and continual public scrutiny of the MR facilities came a need for precise and reliable accounting of facts surrounding any incident which had occurred. RLSAs who were unable to read had a terrible time keeping up with policies and procedures and reference and training guides. Also, pages of progress notes had to be kept by RLSAs and numerous habilitation plans had to be read and carried out. What this meant from a practical standpoint was that each RLSA had to be counted on to follow the policies set by the agency as reflected in the written word and be the person who made the initial report on any incident which adversely affected a client. A case was made that reading and writing seemed to be important part of the job; at least from an administrative and legal standpoint.

Next came job audits. A team of seven specialists from the Office of Personnel Management Personnel Assessment Division made on-site visits to note tasks performed by RLSAs and to do tentative link-ups of underlying knowledges, skills, abilities and other characteristics which aided successful performance on the tasks. Time was shared with RLSAs in various settings at

the MR facilities ranging from those serving heavily involved children who required constant medical supervision to older children not so involved who were learning prevocational skills. The job analysis consisted of both observance of the job being performed and interviews of incumbents and their supervisors.

The job analysis yielded twenty-five tasks which could be agreed upon and cross validated with observations made in other facilities. Sixty-three KSAOs were then linked to tasks and rating booklets formed on which incumbents were to rate tasks and KSAOs on relevance to tasks, criticalness, EOD requirements, and differentiation. The rating booklets were sent to RLSAs in all facilities and their lead persons. Fifty of the rating booklets were returned.

Analysis of Results

In order to organized KSAOs which survived the rating process, criticality ratings were submitted to Principle Components Analysis and rotated to a varimax solution. The minimum Eigen value for the retention of factors was set at 1. The analysis was conducted using SAS.

Five factors emerged which were fairly easily labelled. In order of explained variance, five factors labelled as "Nurturance" (e.g., Ability to care for and remain interested in the well-being and development of clients with few rewards and results for efforts), "Need to Read" (e.g., Ability to act decisively and to react swiftly and effectively in problem situations), "Assertiveness" (e.g., Ability to withstand intense and unexpected displays of affection and aggression), and "Cooperation", (e.g., Tolerance for taking orders and directions from numerous persons) emerged form survivor KSAOs.

The findings led this author to conclude that not only could there be some justification for testing for abuse potential related to the role being a nurturing person plays in performing the job, but there also seemed to be support in the findings for testing for basic literacy and checking on assertiveness, cooperation, and vigilance of applicants within the context of a background check, (in addition to a criminal check). These findings, along with the administrative concern, gave a pretty clear picture of the selection components which could be used in a competitive selection process for direct care workers of the mentally handicapped.

Assessment Elements

Assessment tools were developed after the data analysis just described. A literacy test was developed. One part of the test was directed toward the reading comprehension of materials which closely matched written matter used on the job in both difficulty

level and content. The remainder of the test required a comparative analysis of sentences to pick out those which were the most detailed, clearly stated, etc. , which were closely matched to incident reports which had been completed by persons on the job.

The test has been pretested and has split-half of KR-20 reliabilities in the .90s.

A prototype background investigation form was also developed. The form asks previous employers for references of tan applicant to share examples when the applicant engaged in a behavior or behaviors which could be defined as showing their ability to be vigilant, assertive, and cooperative as defined on the form. The form has yet to be pretested and the scoring guide with an anchored scoring system has yet to be developed.
The most controversial element in the assessment scheme regards screening for child abuse potential.

Child Abuse Potential Inventory

A test purporting to measure child abuse potential in adults was independently discovered by my department manager and myself. My department manager learned of the Child Abuse Potential Inventory (CAP) (Milner, 1977) through a program director at an adolescent diagnostic center. I had learned of the measure when elected to the board of a United Way child abuse prevention program. The author of the test was a consultant for the program and a funded researcher. We inquired into the suitability of the test for an applicant population.

The test is labelled as a "Questionnaire" and lists 160 statements with which examinees must agree or disagree. The statements are written on a fourth grade reading level. Over the years, Milner has developed an abuse scale, a random response scale, a fake-good scale, a fake-bad scale, and others (Milner, 1986). There are over 100 published studies of the use of this measure in predicting abuse in biological and foster care parents. The abuse scale properly classifies known abusers from nonabusers at better than a 90% rate. This rate increases with the use of a "lie" scale. Both backward looking, concurrent, and true predictive validity data are available. Reliability figures are consistently in the .90s for a wide variety of populations (Milner, 1986).

Evidence for construct validity is seen in persons with elevated abuse scores being more likely to report a history of childhood abuse with higher scores reflecting more chronic abuse than the lower scores. Persons with elevated abuse scores have low self-esteem and poor ego-development. Persons with elevated abuse scores also tend to be immature, moody, restless, self-centered, evasive of responsibility, lonely, and frustrated. Mothers rated as nurturing parents have lower abuse scores than the norm

and, of course, far lower scores than mothers known to have abused their children. Though whites and blacks have differences in average abuse scale scores, they are screened out in equal proportions using the author recommended cut-off score (Milner, 1986).

Screening for abuse potential seems feasible due to evidence that abusers have strong involuntary responses and have negative cognitive biases towards children. Frodi and Lamb (1980) demonstrated that known abusers exhibit strong autonomic signs in the presence of children. Pruitt and Erickson (1985) showed childless subjects with high CAP abuse scores to demonstrate the same reactions as in Frodi and Lamb (1980). Twentyman and Plotkin (1982) demonstrated that abusers maintain cognitive distortions in estimating the attainment of developmental milestones of children. Larrance and Twentyman (1983) showed that abusers have negative cognitive biases in terms of casual attributions--they see negative behavior in children as stable and internal while positive behavior was unstable and external. Evidence points to abusers as having characteristics which can be reliably measured.

The bottom line is that there appears to be some technology available which could be of value in assessing direct care workers for child abuse potential. However, numerous issues must be resolved before a measure as this can be used. Here are some of those issues.

Issues in Using Child Abuse Screening Tool

1) The ownership and sequencing problem. Who would actually be designating applicants as having "failed" the abuse potential measure and where will administration of the measure fall in the selection process?'
2) Test score security: Who will be safeguarding abuse potential examination scores and making sure they are kept confidential?;
3) Feedback systems: What CAP data will be released to failing applicants (if any) and how will it be released?;
4) The labeling problem: What can be done to diminish the stigma associated with a failing test score given the title and purpose of the test?;
5) The retesting problem: Will applicants who fail the test be allowed to take the test again as though it were a "standard" merit test?

Future Plans

If these issues can be ironed out, it is the wish of OPM to conduct a concurrent validity study using the CAP to determine if there is any meaningful relationship between test scores and certain administrative and constructed measures used with incum-

bents. If the results are not negative, then a predictive validity study will be initiated using incumbents and new hires who will be given the test, monitored for any abusive behavior, though not screened out with the test. If the results again are not negative, it would seen that the use of the test could be supported for this population of employees and used as part of an assessment scheme.

Selected Bibliography

Atten, D.W. , & Milner, J.S. (1987). Child abuse potential and work satisfaction in day care employees. Child Abuse and Neglect, 11, 117-123

Ayoub, C., Jacewitz, M.M., Gold, R.G., & Milner, J.S. (1983). Assessment of a program's effectiveness in selecting individuals "At Risk" for problems in parenting. Journal of clinical Psychology, 39, 334-339.

Frodi, A.M., & Lamb, M.E. '1980). Child abusers' responses to infant smiles and cries. Child Development, 51, 238-241.

Larrance, D.T., & Twentyman, C.T. (1983). Maternal attributions and child abuse. Journal of Abnormal Psychology, 92, 449-457.

Milner, J.S. (1977). The Child Abuse Potential Inventory: Webster, NC: Psytec Corp.

Milner, J.S. (1986). The Child Abuse Potential Inventory: Manual (2nd ed.). Webster, NC: Psytec Corp.

Milner, J.S., & Atten, D.W. (1985). Institutional child abuse and neglect: A bibliography. Psychological Documents, 15, 22. (ms No. 2715)

Milner, J.S., & Ayoub, C. (1980). Evaluation of "At Risk" parents using the Child Abuse Potential Inventory. Journal of Clinical Psychology, 36, 945-948.

Milner, J.S., & Gold, R.G. (1985). Internal consistency and temporal stability of the Child Abuse Potential Inventory. Psychological Documents, 15, 22. (ms No. 2716)

Milner, J.S., Gold, R.G., Ayoub, C., & Jacewitz, M.M. (1984). Predictive validity of the child Abuse Potential Inventory. Journal of consulting and Clinical Psychology, 52, 879-884.

Milner, J.S., Gold, R.G., & Wimberley, R.C. (1986). PRediction and explanation of child abuse: Cross-validation of the Child Abuse Potential Inventory. Journal of Consulting and Clinical Psychology, 54 865-866.

Milner, J.S. & Robertson, K.R. (1985). Development of a random response scale for the child Abuse Potential Inventory. Journal of Clinical Psychology, 41, 639-643.

Pruitt, D.L. & Erickson, M.T. (1983) A Preliminary Study of a Predictive Model for Child Abuse. Paper presented at the meeting of Southeastern Psychological Association, Atlanta.

Robertson, K.R., & Milner, L.S. (1985). Convergent and Discriminant validity of the Child Abuse Potential Inventory. Journal of Personality Assessment, 49, 86088

Twentyman, C.T., & Plotkin, R.C. (1982). Unrealistic expectations of parents who maltreat their children: An educational deficit that pertains to child development. Journal of Clinical Psychology, 38, 497-503.

* * * * * * * * * * *

"SCREEN-OUT" VS. "SCREEN-IN" TWO MODELS FOR PRE-EMPLOYMENT PSYCHOLOGICAL TESTING

Robin E. Inwald, Director
Hilson Research, Inc.
Kew Gardens, New York

During the past decade, employers have become aware of increasing liabilities attached to hiring "unsuitable" applicants. This has become particularly relevant for those in charge of screening for positions in "high risk" occupations, such as police, fire, or security officers. In recent years, many personnel administrators have turned to psychologists for assistance in making their hiring decisions. Where there is potential for "negligent hiring" lawsuits, psychologists have been called upon to aid in the detection of emotional instability and/or disorders that could result in serious difficulties on the job.

Psychologists have responded to the needs of administrators by providing batteries of psychological tests and clinical interviews, often used to document reasons why an individual should not be hired. The "medical model" has been favored, which looks for clinical abnormalities and psychopathology in applicants. While the MMPI remains the most common instrument used in the effort to "screen out" individuals with "problems," several newer instruments have also been developed to detect behavior patterns and attitudes that are predictive of poor performance.

One test, the Inwald Personality Inventory (IPI), has been used in hundreds of police, correction, and security agencies to aid in the prediction of absence, lateness, and subsequent termination of officers. This instrument includes scales such as Alcohol Use, Drug Use, Trouble with the Law, Job Difficulties, Absence Abuse and Interpersonal Difficulties. Such scales are behavioral in nature and focus on "negative" past behaviors as a key to predicting future job adjustment difficulties.

Research on written tests has indicated that some utility is gained by using tests of "negative" behavior. One study, a five-year follow-up of over 200 officers to be published in the November, 1988 issue of the Journal of Applied Psychology, reports "hit" and "miss" rates associated with various "cut-off" scores. In this study (Inwald, 1988), it can be seen that specialized IPI prediction equations can identify roughly half of those who will be terminated within five years, while falsely predicting 11% to fail who will not. While IPI and MMPI prediction equations based on scale scores alone can identify up to 69% of the true "failures", they result in false positive rates of over 26% and 35% respectively.

Another method for screening prospective employees involves focus on "positive" attributes. While drug use and other clearly "negative" behaviors may not be detected using a "positive" screening method, the benefits are that this kind of screening may help employers discover talents in applicants that can lead to development of abilities and future promotions. With limited training resources, it is increasingly important to place new employees in positions that can capitalize on their strengths and will not be adversely affected by their shortcomings.

The Hilson Personnel Profile (HPP) was developed in an attempt to identify some universal qualities most critical for "success". Scales focus on behavior patterns and styles found in successful individuals in their fields. The HPP consists of 150 true-false items grouped into five major scales: Achievement History (33 items), Social Ability (40 items), "Winners" Image (28 items), Initiative (33 items) and Candor (16 items). Three of the HPP scales contain items that have been divided into separate "Content Areas". These include Social Ability: Extroversion, Popularity, Sensitivity; "Winners" Image: Competitive Spirit, Self-Worth, Family Achievement Expectations; Initiative: Drive, Preparation Style, Goal Orientation, and Anxiety about Organization.

Over 900 entry-level job applicants were administered the HPP along with over 300 working individuals, including professionals and entrepreneurs. The average alpha coefficient for entry-level applicants was .70 and the average for employees was .81. These results suggest that each of the five HPP scales are internally consistent and reliable. A factor analysis revealed a single

factor for the job applicants, while 455 working individuals from various organizations showed two factors. The first factor included Achievement History, Social Ability, "Winner's" Image, and Initiative, while the second included high Candor and low Initiative only. This second factor appeared to include individuals who know themselves well, are satisfied with their careers, and who are not particularly "driven" to excel in their fields.

When HPP scales were correlated with the MMPI and IPI, they showed few correlations greater than .29. "Winner's" Image negatively with Hysteria, and Social Ability correlated negatively with Social Introversion on the MMPI. However, with so few correlations between the "screen-in" and "screen-out" tests, it can only be said that the absence of negative behaviors/psychopathology does not mean the presence of "positive" work adjustment.

Finally, when the HPP was used to identify exceptional employees in a number of companies, it was observed that, in general, the more scores higher than 59t, the more likely the individual was to have received a positive rating by his/her supervisor. Much future research is warranted in order to develop the HPP for use in predicting future positive job performance in different occupational categories. However, these data suggest that a two-pronged approach using both "positive" and "negative" screening instruments may provide different, but equally helpful, sources of information for hiring decisions.

* * * * * * * * * * *

EXAMINATION SECURITY -

HIGH TECH OR LOW TECH?

Lee Mattice
Assistant Director of Evaluation
Michigan Department of Civil Service

We, as test administrators, have the responsibility for developing and maintaining a security plan or system to assure the integrity of our product. The product referred to here is the examination. The security plan or system should not be developed as a reaction to a problem but should be a specific plan with established objectives. These plans must be put in place and strictly followed to protect against those individuals who would profit from their ability to breach our security.

Some articles of security breaches are include at the conclusion of this paper. As you will note, after reading these articles, various methods have been used, including attempted bribery, theft of examination materials, and using a substitute to take the examination, in order to gain a competitive advantage in the examination process, or for some other form of gain. When you read these articles you will probably recall similar events that have happened in your jurisdiction or to someone you know.

For years the Michigan Department of Civil Service thought its security measures were adequate. We were satisfied that our processes, a computerized item file, limited access to examination materials by staff, examination material auditing, and trained monitors constituted a strong security plan. We were satisfied that our processes prevented the possibility of security breaches, and if any occurred, we would immediately respond and take the appropriate action.

However, the unsolved theft of a promotional examination booklet and its impact on subsequent test scores proved that our system and our reactions were insufficient.

As a result, the Department initiated two actions. First, all facts and information regarding the missing booklet were collected and turned over to the Michigan State Police for investigation. Second, an Examination Security Committee within the Bureau of Selection was established. Part of the committee's function was "...to review the Department of Civil Service's examination security process and identify weaknesses and recommend suggestions for strengthening the process."

At the conclusion of their review, the Committee presented a number of recommendations. Some of these recommendations with a brief discussion follow.

Develop and adopt an examination security rule.

During both the Committee's review and the State Police investigation it was determined that there was no Civil Service Rule, nor any State statute protecting Civil Service examinations or State of Michigan licensing examinations. Without such a rule or law, there is no legal protection covering the examination.

Develop and adopt an explicit examination security plan.

Our current plan is in pieces, covered in administrative rules and in internal operating statements in various sections within our bureau. It is our intention to bring all the pieces together, with any additions, to develop a comprehensive security plan.

Establish an on-going audit team.

The Bureau of Selection Director, at his discretion, will periodically use staff from this and other bureaus to compare the security plan with actual practice to assure that the plan is being followed.

Additionally, staff from our office will visit the examination centers, on a periodic basis, to assure that all monitoring and security functions are being followed at the centers.

Provide a security work station.

The use of open-space work stations has reduced our ability to maintain security in an individual office. It is our intention to redesign our examination security room to incorporate several security work stations. The redesign also includes replacing the standard key locks on doors with the security combination style locks. These locks will also be of the type that will allow resetting the combinations periodically.

Use security agreements for subject matter experts, Civil Service staff, and others.

This agreement outlines and defines the role and responsibilities of persons relative to contact with examination materials from test development to general security of all examination material.

In addition to the above recommendations, the Michigan Department of Civil Service is also implementing or reviewing new security measures. These include the following:

- Developing training videos to be used when new monitors are hired. Content will include definition of the monitor's role and responsibilities, security measures, and observing applicants for possible cheating or collusion.

- Using scrambled forms of the test booklet. The original and the scrambled version will be alternately distributed to applicants to discourage copying.

- Using numbered test booklets for additional control. A missing booklet can be traced to a person, or between two persons.

- Using new wrapping methods when shipping examination materials. Instead of using wrapping paper or tape, use shrink wrapping.

- Using locking cases when shipping examination materials. Cases will have combination locks that can be reset on a periodic basis.

- Collecting and holding the applicant's I.D. when the test material is distributed. The I.D. is returned to the applicant when the test materials are returned to the monitor.

- Limiting access to examination materials and to the automated item file to only those individuals who work with the materials. This is done not because of staff is not to be trusted, but to protect staff if there is a breach.

- Providing locking file cabinets within the security room for all confidential examination material. Only the security room attendant and the attendants's supervisor have keys to the cabinet. Any confidential material being removed from the security room must be signed out.

- Providing alarm systems and television surveillance of the security room to guard against after hour or unauthorized entry.

Monitors are instructed to follow procedures at the examination centers. These include:

- Remaining in the examination room. During the State Police investigation it was determined that monitors were leaving applicants and materials unattended in the examination rooms.

- Observing the applicants. It was also reported that monitors were congregating in corners, or at the main desk, talking, reading newspapers, or performing tasks other than that of monitoring.

- Looking for various methods of cheating. Applicants used slips of paper the fit in the palm of the hand. This slip of paper had the answer key. Also, other unauthorized aids were used.

- Securing the unused examination material, before, during, and after the test to assure that copies cannot be made.

- Assuring that only test related materials are on the testing surface. All other materials are to be kept off the testing surface.

- Collecting all materials, including scrap paper, used during the test session.

The above mentioned security measures being taken indicate the value we place on our product. It may appear that the steps outlined above are excessive or expensive. However, when you compare the cost of replacing the test, and the loss of credibility, I think we would all agree it is money well invested.

The security measures listed above, by no means, are intended to be all inclusive. We will continue to monitor our progress and adopt whatever security measures are required to protect our examinations.

* * * * * * * * * *

THOUGHTS ON EXAMINATION SECURITY

Thomas A. Tyler
Merit Employment Assessment Services, Inc.

One security problem that every agency shares is the security of records after the tests. It is not unusual for important documents, such as eligible lists, to mysteriously disappear. Needless to say, it always seems to serve someone's purpose when this happens but it is almost always impossible to recover. One simple solution to this problem is to file all documents of this type with the appropriate governmental filing office. In Illinois this is usually the County Recorder of Deeds but in very small counties it may be the County Clerk or Registrar. Once a document is filed in this manner it is available to any interested party willing to pay the small copy fee. The Recorder of Deeds usually makes a second microfiche for the document which it stores in the State Archives. Retrieval of the document is easiest with the document number, but searches can be made for any document.

Occasionally, you might have a document that is absolutely top-secret, perhaps even from you. Suppose, for example, you want to collect performance data for a validity study but the raters are reluctant to make such ratings because, in a previous situation, their ratings were subpoenaed and made public in a court case. In this case find a Canadian colleague; have the ratings mailed directly to your colleague. Have your colleague code that data by an anonymous ID number and return the coded data to you. Furthermore, instruct your colleague to keep all of the information secret, even from you and even if you ask for it. This procedure should keep the sensitive information safe, even from a subpoena.

Even if you do not intend to use your tests a second time it is a good idea to copyright them. It only costs $10.00 for a copyright, and you need not file the required two originals until after the test has been administered (within five years). The advantage of a copyright is, that again, you have a permanent record someplace, and that you gain some control over what happens to your test should a copy fall into the wrong hands. For example, a copyright will discourage the local newspaper from printing the test ---- to your embarrassment.

For high security tests the U.S. Copyright Office provides a system for filing a "mutilated" unreadable version of your tests, that maintains a sufficient identification (with the unmutilated version in your possession) to provide copyright protection.

Now for some odds and ends: Use colored paper for your test covers. The colors allow quick visual cues during test administration and can help you spot a test booklet that is not where it is supposed to be.

Occasionally I collect a thumb-print right on the answer sheet as the test is handed in. Sometimes the police identification section does this for me and sometimes I just use an office stamp pad. This procedure discourages "ringers" even though my stamp pad impressions would probably not hold up in court.

Often, the illusion of security is as important as real security. Make a big show of your security procedures. If your test administration is too crowded alternate the colors of the test booklet covers. Even if the same tests are between the covers, the candidates will believe they have been given different forms.

Impress your candidates with your professional status. When you introduce yourself say, "My name is John Smith. I am a member of the Assessment Council of the International Personnel Management Association and am bound by the professional ethics of that Association, etc." The more you can make the candidates believe the exam is being done professionally, the better your security will be. Dress the part too; you need to look like an authority figure.

You can contribute to the illusion of security by ushering candidates to the washroom and usher candidates from their seats to the place of exit. Put an official-looking seal on the edge of your test booklets. Use those transparent envelopes from 20th Century Plastics to bundle your booklet, ID set and answer sheet. They are reusable. Never, never, never allow candidates to stand up and leave their seats at the end of a test. This makes a crowd around your exit table and that is where you lose test booklets.

Count your tests when you get them, when you lock them up, whenever they change hands, before you give the test, during the administration, immediately after the test, and when you destroy them or return them to your publisher. If you do lose a test booklet you should know when you lost it. That is important.

Remember, the carbon ribbon out of your typewriter is almost as good as a photocopy and your print shop is likely to discard spoils and plates in the common garbage if you are not there to watch them.

If you maintain a file of tests that may be used again, in whole or in part, have manila envelopes preprinted with an inventory control form. This form should indicate the date and use of the test, how many copies of the test, how many copies of the test key, etc., are in the file. Have a "check-out" card made to be placed in the file when any item has been removed by your staff.

If you have a candidate with questionable eligibility show up to take the test, always allow that candidate to take the test. It is far easier to disqualify the candidate later than it is to maintain security for a second administration.

Take a box of kleenex to your test site. Some candidate always has a cold and no handkerchief -- you will save one washroom escort. In larger test administrations it is likely you will have someone sick to their stomach. Consider a mop and pail and janitorial service.

Good test security means planning ahead and being prepared for most contingencies. The better you plan ahead the more relaxed you will be and the better you will be able to cope with the unexpected.

* * * * * * * * * * *

THE "LOW TECH" OF TEST SECURITY

Barbara Showers, Director
Office of Examinations
Wisconsin Department of Regulation and Licensing

Test security is not taught in college, even in testing and measurement curricula. It is developed through experience. Some people have a talent for this. They may tend to be authoritarian and picky. Try to hire them in positions responsible for test security. Good sources of information on test security measures can be found in the test administration manuals of large testing companies. Many jurisdictions have also developed manuals on

this topic. Providers of licensing examinations are particularly vocal on the topic of security. The contract for purchase of the nursing examinations contains 27 pages of security measures which must be followed.

Elements of security are control, traceability, verification, and responsibility. I will attempt to highlight some key points of routine test security that must be considered in the business of test development and administration.

Pre- and post-administration

a. While developing the test consider:
   1. Office desks in low public access areas, and vault storage of all files.

   2 Limited access to computer files and work processing or typist procedures, including intraoffice delivery of documents.

b. While printing the test consider:
   1. Supervised printing
   2. Return waste with printed copies and destroy.

c. While storing and delivering the test consider:
   1. Need for inventory audit trail to track when and where the booklet became missing.
      a. Number the booklets
      b. Inventory when packing, before giving at site, after giving at site, on return to storage.
      c. Provide physical barriers such as string or shrink wrap, and ideally, sealed booklets which show evidence of tampering.
   2. Use a traceable method of delivery (UPS, Air Freight)
      a. Specifically inside delivery to a person who will be there when delivered.
      b. Pack tightly in sturdy boxes so they don't spli' open.
      c. Don't advertise the content of the boxes if possible.
   3. Provide limited access storage at the site and the office
      a. Who has the keys? Often maintenance staff.
      b. Use key core or key block at site.

On Site Administration

a. Most common types of cheating to control: taking the booklets, looking on another's paper, hidden notes--having or taking out, and impersonation. Others (handout).

b. Control measures:
   1. Admission and seating: admission tickets, photo identification, seating charts and preassigned seats, and

spacing of seats (5 feet on either side, examples of seating plans.)

2. Movement: Single entrance and exit, permission to leave, restroom monitors, avoid crowding the checkout.

3. Control of booklets: Pass booklet directly to candidate, no unattended piles of unused or collected, booklet collection point away from exit.

4. Proctors:
   a. Well trained and qualified. Signed security contract
   b. Specific responsibilities (Overhead list).
   c. Sufficient numbers (At least 1 to 35. May need more if site has poor layout or complicated ad ministration.)
   e. Recommended action (handout, discussion)

Finally, be sure to have plenty of evidence before withholding a score due to cheating.

1. Accurate observation by multiple people and writeup.
2. Comparison of answer sheets if copying.
3. Physical evidence, e.g., notes if possible.

While much of test security is "low tech" and administrative, it requires considerable commitment to maintain. A representative of a large testing company recently stated the belief that most cheating goes on undetected, especially in the areas of impersonation and copying.

Test security is a "low tech" area that requires high priority in the management of a quality testing program.

* * * * * * * * * * *

DEVELOPMENT AND IMPLEMENTATION OF AN INTERACTIVE ORAL EXAMINATION FOR JUVENILE CORRECTIONAL WORKER

Nancy J. Skilling
Hennepin County Personnel Research,
Minneapolis, Minnesota

Knowledges, Skills, Abilities & Personality Characteristics

- Knowledge of Adolescent Development
- Knowledge of Group Dynamics

- Knowledge of Juvenile Delinquency
- Knowledge of Learning Theories
- Knowledge of Counseling Theories
- Knowledge of Chemical Dependency
- Knowledge of Normal/Deviant Behavior
- Knowledge of Corrections
- Knowledge of Human Sexuality
- Skill at Oral Communication
- Interpersonal Skills
- Judgment/Decision Making Skills
- Problem Solving Skills
- Ability to work as a Team Member

Goals

- Free From Adverse Impact
- Test Critical KSAPs
- Interactive Format
- Applicant Friendly
- Readily Scored

Development Steps

- Review previous job analysis data
- Review critical KSAPs with SMEs
- Review and modify previous oral exam items with SMEs
- Develop new oral exam items with SMEs
- Develop response guidelines with SMEs
- Develop exam and training materials
- Train oral board members
- Administer oral exam
- Analyze oral exam results
- Conduct feedback sessions with hiring department

Each Situational Item was Measured on:

Action Scale:

What the Candidate Indicates They Would Do

Rationale Scale:

Why the Candidate Would take these actions

Situational Items For Oral Examination

1. Conflict over a Dinner Rule

2. Ramone's Refusal to Work: Prior to Escalation
   Ramone's Refusal to Work: After Escalation

3. The Racial Joke

4. Terry and You

5. Joel's Refusal to Go to School

6. Teresa and the Suspected Drugs

7. Darren's 120-Day Stay

Other Ratings

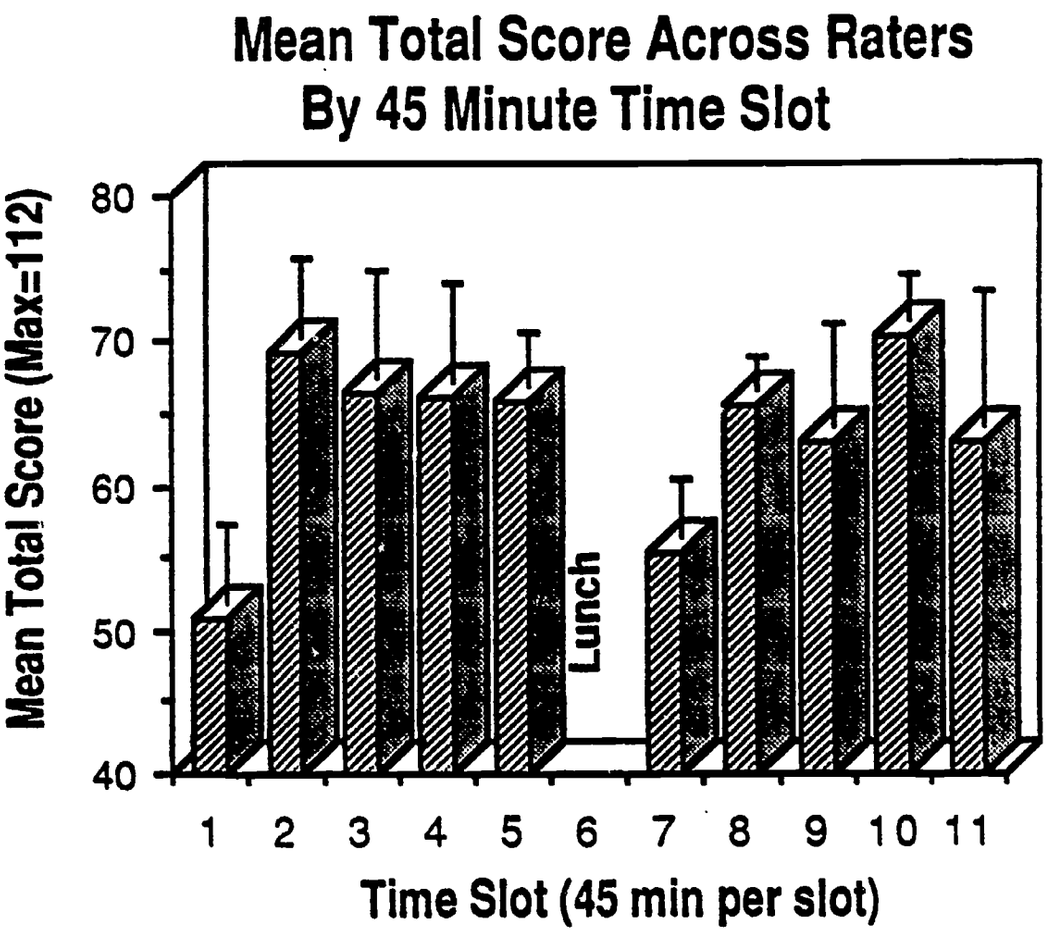
Overall Suitability to Perform as Juvenile Worker

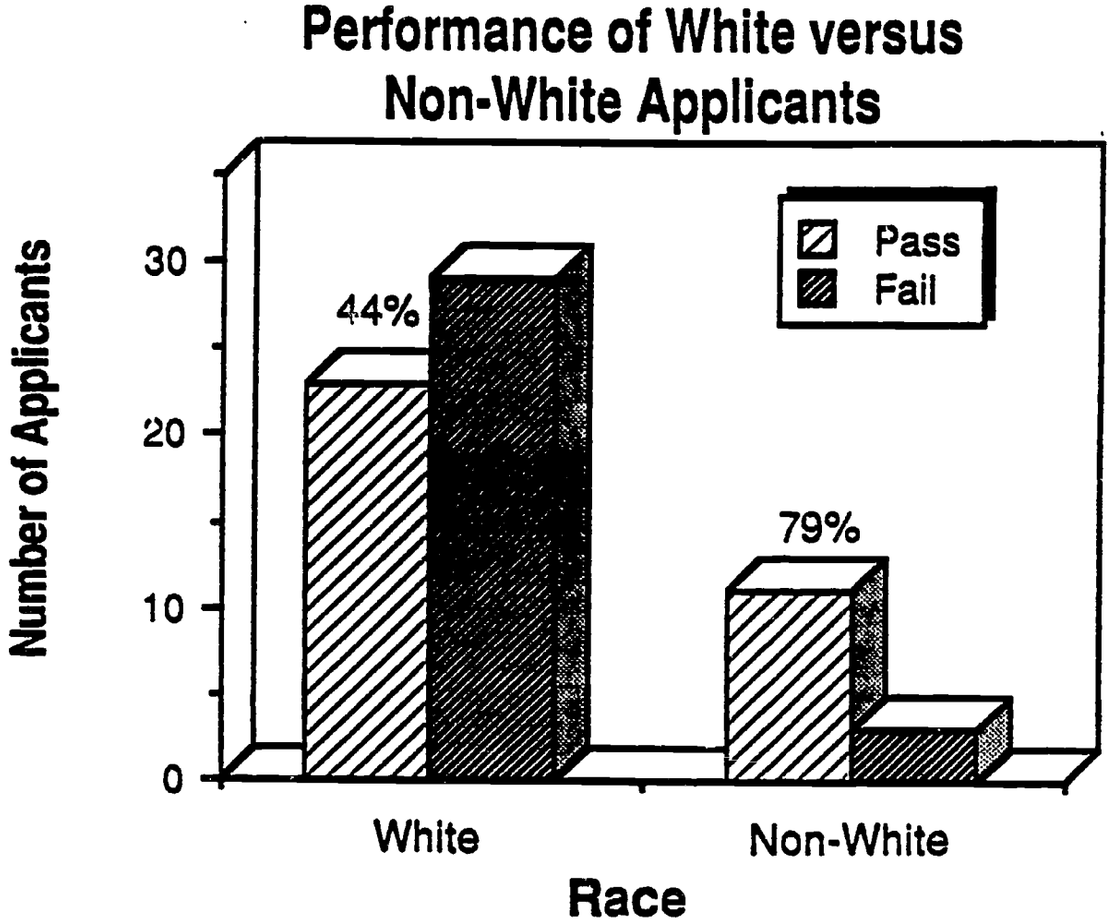
Confidence in Ratings

Would you hire this person?

Is this person trainable?

Could you supervise this person?





* * * * * * * * * * *

A COMPARISON OF THE ORAL INTERVIEW AND BEHAVIORAL CONSISTENCY EVALUATION METHODS FOR SELECTING JOB APPLICANTS

Sally A. McAttee, Director of Examinations
City of Milwaukee, Wisconsin

This study compared the oral and behavioral consistency examination methods in the selection process for two managerial positions. The need for such a study arose from the researcher's

desire to find a testing method which possessed the desirable characteristics of the oral interview but which avoided its disadvantages. The behavioral consistency approach was used as an alternative to the oral interview because it is parallel in development, content, and administration but involves no interaction between raters and candidates.

For each position, test development for both approaches was based on a job analysis which defined the essential job dimensions. Test content was parallel. The behavioral consistency examination asked candidates to describe major achievements which demonstrated their capabilities in each job dimension. The oral examination consisted of two questions developed by subject matter experts for each job dimension. There were 18 subjects in the first sample and 14 in the second.

The findings were as follows:

1. The results regarding the comparability of the two methods were inconclusive. Correlations between the methods were significant and meaningful for one sample but were non-significant for the other.
2. There were no significant differences in reliability between the two methods for either the overall ratings or the dimension ratings for either sample with one exception for the dimension ratings.
3. Convergent validity results were inconclusive. The methods demonstrated convergent validity for one sample but not for the other. The methods did not demonstrate discriminant validity for either sample.
4. There were no significant differences between the methods regarding their acceptability to the raters. However, based on descriptive comparisons, the behavioral consistency method was superior in terms of rater time.
5. Based on descriptive comparisons, time efficiency for the candidate was in favor of the oral examination. However, candidate time included only actual examination time; it did not include time for travel or preparation.

* * * * * * * * * * *

TYPES OF MULTIPLE-CHOICE QUESTIONS

THAT MALFUNCTION

Chuck Schultz and Brenda Morefield
Washington State Department of Personnel

We often assume that when a candidate fails a test he or she lacks the quality the test is supposed to measure. The variables deliberately measured by employment tests are the knowledges, skills and abilities related to superior job performance. Our test development procedures ensure our tests measure these. The subject-matter specialists verify that we should be asking the kind of questions in the test.

But factors besides what the test is intended to measure affect test scores. Because of these other factors, people who know how to handle a situation may not give the "correct" answer to a question about it on the test. Candidates wonder in what frame of reference to respond. They must decide whether to state a solution, obtain more information, or refer the case to someone else.

Over the years we have identified many types of test questions that have not worked as intended. Certain question formats result in candidate response patterns that cannot be explained in terms of question content. The formats seem to elicit responses that are more related to "response sets" than to an understanding of the subject matter. Different candidates' expectations about the test lead to different response patterns.

Let's look at some question formats that lead to malfunctioning questions and discuss how to improve them.

Negative wording. We used to pose questions in the negative. We might ask, "which of the following is not a factor in..." or "which of the following is least important to...". These produce peculiar results. The candidate may understand the question initially, but, in the process of analysis, the candidate concentrates on the issues and forgets the negative orientation.

Question 4 is another kind of numerical question that shares the one-smaller-one larger bias. If I don't know how to solve for the area of a triangle, I can figure out that the area cannot be more than half the product of the two shorter sides. Therefore, I'll pick 54 as the better choice of b and c.

We can make numerical questions more fair by giving a and d equal time. "When in doubt pick b or c will no longer give the test-wise an advantage. Then all candidates have only one chance in four of stumbling into the correct answer.

We have added a fifth alternative, e. In numerical questions we include something like "some other amount" to alleviate another artifact. Without it, candidates who make the biggest mistakes get a second chance.

Since distractors in quantitative questions are designed to be the most likely wrong answers, those who make reasonable mistakes mark one of the alternatives offered. People who make unreasonable mistakes won't find their answers among the distractors, so they have to try again. Therefore, the person who makes the worst mistake gets a second chance, while the person who makes a common mistake happily chooses one of the distractors we provided, and misses the question.

"Some other amount" fits any outlandish solution. We use it as the keyed answer one time in five to neutralize the test-wise.

True-false. True-false questions are sometimes place in a quasi-multiple-choice form by asking something like, "How may of the following statements are true?" We do not like the connotation of absolutes implied in true-false questions. A statement has to be blatant to be false in every conceivable situation. Candidates differ in judging how true a statement must be to be called true.

Take for example the true-false questions 5 through 8. You can make a case that any one of these is true. You can also make a case for any one's being false. Question five: There are other considerations than utilization of staff for assigning tasks, so 5 can be false. Question six: Wile employee preferences should be considered, the organization's mission is more important, so 6 can be false. Statements seven and eight are contradictory, so if one is considered true the other could be considered false.

On questions such as these, whether a person answers true or false depends on more than the person's understanding of the issues. It depends on how one interprets the situations. Questions like this sometimes appear on an objective test, but how objective are they?

We do not use all-of-the-above questions.

Social Desirability. The social desirability response set has been studied extensively in personality tests. Social desirability is active in multiple-choice tests as well. All too often the correct response is clearly the most socially acceptable thing to do. In questions 11 through 13, we leave off the item stems and present only the alternatives. As you read through those alternatives you will probably see that some of the actions are quite socially desirable. The numbers in the left hand column show how many candidates picked each alternative. We had data for 130 candidates for 11, 12, and 13.

Almost all the candidates chose the socially desirable response in 11, 12, and 13. On question 11, there are some candidates who believe in confrontation as well.

For those questions the socially desirable response was the keyed answer. The problem we see here is that you don't have to respond to the question itself--you don't even need the stem to get the question correct.

In multiple-choice test questions, we have seen that often the keyed response differs from one of the distractors only by the social acceptability of the phrasing. Question 14 is such a question. Alternatives a and b are two different ways of saying "do nothing", one of which is more attractive than the other. Alternatives c and d are two ways of saying "ask her to be nice". This is a two-choice question. Candidates will select either b or d.

On questions 15 and 16 the problem takes on a different hue. The subject-matter specialists told us that Caseworkers need to know when to close the case. They created situations in which you have done everything you are supposed to do for the client and there is nothing more you legally can do. So you are supposed to close the case. The keyed answer for 15 is a and for 16 it is b. However, at least on the test, candidates find a variety of services that are preferable, more socially desirable. If they have the option of saying they would close the case or saying they would do something more friendly, candidates pick the more friendly answer.

Are the few people who said they would close the case the best candidates? That is not clearly so. The social desirability response set seems to be working against us.

Using the same words in every test booklet does not ensure that all candidates have the same question. The words mean something different to each of us.

We may want to see whether the candidate knows that acting without more information is premature. We expect the candidate to know that this time more information is needed. Other times we fail to provide all the information one would have on the job and expect the candidate to extrapolate. How can the candidate tell which is the case on a particular question?

Look at question 18. We have given you some information about Carl and his family. Do you have enough information or should you gather more before charting a course of action? We find that some good caseworkers choose to solve the problem on what we have given them, while others feel they need to have more information. To put all candidates on the same wave length, make the alternatives parallel.

Frame the question one way or the other. If you want to find out about the candidate's ability to determine what information is needed, ask about information, as we do in question 19. If you want the candidate to make a decision with the information at hand, give as alternatives various courses of action, as we do in question 20.

What if the worker should do nothing? In some situations one should wait until the time is ripe, but the candidate expects, "They must want me to do something now, or they wouldn't have included this question in the test."

A subject-matter specialist told us, "In this job it is important for the incumbent to be patient." So we wrote a "be-patient" question something like question 21. Since you know our rationale, perhaps you will accept d as the correct answer. But, empirically, the question did not work. Candidates, even the better candidates, came up with creative solutions. They may act like stodgy bureaucrats once they are in the job, but on the test they are proactive and innovative.

Question 22 may be a better way to see whether a manager will expend resources on a program enhancement that has not been funded.

A question asked how the candidate should handle a situation. For a Contracts Specialist 2, the appropriate response was to notify a higher authority. Instead these excellent candidates told how the situation should be handled. Shame on them! Or shame on the test writers?

Question 23 is an example of a multiple-choice item that forces the candidate to choose between solving the problem and referring the case to the proper authority. I believe the answer is b, but many candidates may think we want them to do something positive rather than pass the buck. Again, we should make the alternatives parallel, and either give four ways to solve the problem, or four ways to get someone else to handle it. Question 24 presents alternatives at a Clerk Typist's level of involvement. Question 25 deals with how to handle the problem, but it is not directed to a Clerk Typist.

Multiple-choice questions need to be stated in such a way that each candidate will be able to see the level on which the question should be answered. To this end, the alternatives should be parallel. Should the candidate collect more data, refer the case to someone else, or close the case rather than solving the problem? Make it clear whether the candidate should select a solution to the problem or a way of dealing with the case preparatory to formulating a solution.

In conclusion: we need to ensure that the test score results from intended content not from artifacts. Don't use faulty formats. Use parallel alternatives. Phrase the questions so that candidates know what tasks to address.

* * * * * * * * * *

PROBLEMS OF BIAS AND TEST-WISENESS IN MEASURING ORAL COMMUNICATION AND PROBLEM-SOLVING SKILLS THROUGH MULTIPLE-CHOICE ITEMS

Christina L. Valadez
Washington State Department of Personnel

Although varying in degree of importance from job to job, good communication skills are identified as an important element in almost every job analysis we conduct. Several aspects of communication skills are not typically considered when developing multiple choice items, and yet have the potential of affecting test results.

The test writer faces a challenge in testing for these skills. First is to get the subject matter specialists to define what constitutes good communication for their jobs, and next to determine how the best measure their definition of "good communication." This is particularly challenging when part of the communications skills needed are oral communication skills, yet the testing format is to be multiple-choice. We face this dilemma when we need to conduct continuous or frequent testing for large numbers of candidates in different geographic areas.

The solution we typically rely on is to present a situational problem involving verbal interaction, and ask the candidate how to best solve this problem. A number of verbal strategies are offered as alternatives, and candidates are asked to choose the one they believe is the best response to the situation described. This approach typically assumes some measure of problem-solving ability, another element prevalent in most job analyses, as well as "oral communication skills." Depending on the level and nature of the jobs, other elements, such as "interpersonal skills," "dealing with the public" or "supervision" may be part of such a situational item. The essence of such items, however,

remains "how to effectively respond to a communication problem in a given context."

In an effort to anchor the responses the subject matter specialists identify as important to on-the-job performance, we ask them for behavioral examples. How does your best performer respond? How does a poor performer respond? We use their answers to those questions to build our keys and distractors.

But are subject matter specialists really providing us observations of successful oral communication strategies or are they instead providing us examples of their own style, or perhaps their assumptions of a strategy they believe produces the desired outcome?

It is interesting and instructive to compare how we attempt to measure oral communication skills in a multiple-choice format with how we measure more quantifiable skills, math for example. When we present a math problem to solve, we focus on the end result. Any given problem may allow numerous ways to work out the answer. My personal observations have shown differences in the process used across generations due to changes in teaching methods, and also differences due to the various teaching methods in different countries. The validity of different approaches is recognized through testing for the ability to correctly reach the final result, rather than testing for knowledge of a particular process.

However, when using multiple-choice testing for oral communication skills, by anchoring responses to behavior subject matter specialists proclaim "best", we are measuring the knowledge of the process rather than the ability to attain a successful outcome. It is this assumption of the superiority of one approach in producing the desired outcome that may present problems due to differences in socio-cultural orientation, or due to test-wiseness.

Problems can occur when relying on an organization's verbal behavioral norms for keying a particular aspect of the communication process as "correct." Besides individual and organizational differences in communication style, socialization in what constitutes appropriate communicative behavior varies across ethnic, gender, geographic, and socioeconomic lines.

An example comes form Patricia Clancy's article, "The Acquisition of Communication Style in Japanese" (1986). She documents the efforts of Japanese mothers to teach their children how to express themselves, particularly their desires, in an indirect manner, and how to interpret the indirect requests of others. This focus on indirect expression contrasts sharply with the expressive values of directness found in may of our test items. Test-wiseness or other awareness of norms calling for directness

will lead candidate to the ky, regardless of whether the candidate believes in or uses directness as the "better" strategy.

Testing for the verbal behavioral processes then, rather than for the final outcome or for considerations for reaching the final outcome, may therefore have the effect of testing for verbal socialization patterns. How closely the applicant's patterns of verbal interaction conform to the ideals of a certain organization is likely to be reflected in the test score. This is not the same as testing a candidate's ability to communicate orally in the way necessary to do the job well.

How can we test for this ability without unnecessarily excluding good candidates? Oral communication is a complex web of vocabulary, grammar, structure of narrative, nonverbal cues, social cues, and paralinguistic speech features such as accent, use of "fillers" (hm, uh), rhythm and speed, etc. All of these features combine and interact. Two speakers of the same background share these features and therefore are likely to derive the same meaning from an interchange. There is no doubt that differences in the interpretation of these features can increase the potential of miscommunication.

One of the most interesting features of human communication is not the knowledge of a particular set of rules; but the ability to learn and adapt. Those who are skilled in the art of oral communication can use communicative differences and resulting miscommunication as a source of expanding their understanding, and can adapt to new interactions.

We adapt daily to different modes of communication between work and home environments; between co-workers and the public. Every time we move into new social environments, we begin to learn new ways of interacting with others. How well or how quickly this is accomplished varies from individual to individual. It is this variability that is a truer measure of oral communication skills than knowledge of a preferred communication model. Do current multiple-choice items presumed to measure good communication skills test for this variablilty? I strongly suspect that most do not.

We frequently receive comments from candidates that depending on circumstances which we have not addressed in the multiple-choice item stem, they could choose any of the distractors offered as the best response. We tell candidates to rely solely on the information provided to choose the best response. Yet there is so much paralinguistic information (e.g., tone, volume, word spacing, etc.) and nonverbal information (stance, gestures) not to mention social information (individual history, rank relationships) that we take into account both consciously and unconsciously. Indeed, training in management and communication encourages us to consider numerous factors in communicating with

different individuals differently, rather that always using what falls into our own communicative comfort zone. After much reflection, I am inclined to agree with the candidates who say "it all depends."

Furthermore, my experiences with the subject matter specialists on whose information we rely is they (a) are not always the superior workers we ask agencies to send us and (b) even when they are, they are very conscious of the expectations of their supervisors and organizations as well as what they imagine to be those of the central personnel agency with whom they are working on the exam development. Despite our intensive efforts to sort the wheat from the chaff in job analysis and item development and review sessions, few SMS groups will approve a keyed answer that is contrary to organizational norm, whether or not it is actually what occurs.

Communication is not measurable in the same way typing performance is. The communicative approaches that produce problems are by far more evident than those that are successful. Therefore, successful strategies may be assumed to follow normative patterns whether they do in reality or not.

I'd like to close with some questions to consider as we seek new and better ways to construct our tests to truly measure what we intend. Given some to the problems outlined, how valid is it to test for current knowledge of the norms of appropriate verbal behavior in a particular environment? Even if we argue that current superior workers conform to the organization's communication style or values, how job related is a reliance on one approach, given the diversity of the ever-changing modern workforce? What are other alternative? If we need to rely on a multiple-choice format, how can we better test for true communicative abilities?

In oral exams, we can be much more flexible about crediting a variety of approaches that will achieve that desired outcome. In multiple-choice tests with only one allowable "correct" response, testing for these skills is much more problematic.

Perhaps we need to focus multiple-choice items more on the criteria for achieving desired outcome of communicative problems rather that a "correct" process. And, of utmost importance, we need to make sure SMS description of communication problem solving goes beyond their perceived reality based on norms, to the factual observation that we request of them.

I hope through these means we can develop multiple-choice items that will work better to select the best candidates from a diversity of backgrounds and avoid the test-wise who simply know the rules.

* * * * * * * * * *

IRRELEVANT RELIABLE VARIANCE

CHUC SCHULTZ
TEST DEVELOPMENT MANAGER
WASHINGTON STATE DEPARTMENT OF PERSONNEL

Christina and I have been talking about the test-item characteristics that affect variance. Method variance, test-wiseness, and cultural bias are unwanted sources of variance, irrelevant to the purposes of the test. Variance is the most useful statistical indicator of the amount examinees' scores are spread by different variables. I want to distinguish among irrelevant, relevant, and random variance and how the different components of variance affect test reliability and validity.

First, I conclude that the more irrelevant test variance you have, the higher the reliability. Anything that increases total variance relative to random error increases reliability. Reliability tells how consistently the test measures whatever it measures.

Second, I conclude that having the same biases present in the test and the criterion measure inflates the validity coefficient. If correlated biases are present, the same thing will happen. For example, if for some erroneous reason a rater thinks group A members can't do the job well, and for another erroneous reason group A members do poorly on the test, you have correlated biases. Two wrongs make an enhanced validity coefficient. I emphasize validity coefficient as only one indication to test validity.

A validity coefficient is the correlation between a test and a criterion MEASURE. The criterion measure may or may not be an adequate reflection of the criterion (for example, job performance). You may use any of a number of criterion measures in a validity study, each of which measures something different. You could use measures as diverse as number of units produced, supervisory ratings, or attendance. Each criterion measure gives you a different validity coefficient. The criterion measures probably overlap with one another and each probably overlaps with the hypothetical real criterion. Let me illustrate the relation between variance components and reliability and validity.

Handout 1 pictures the components of variance in a test, a criterion measure, and a hypothetical pure criterion. Let's say you built a test to predict job performance. You designed the criterion measure to check the validity of the test. The criterion itself is a hypothetical construct -- it is the

quality of job performance that we are trying to measure with the test and the criterion measure.

In the handout, the dotted circle stands for this "real" criterion, the heavy circle is the test, and the light solid circle is the criterion measure. Different variance components are represented by the number segments of the diagram. Segments 1, 4, 5, and 6 fall within the dotted circle representing the real criterion. These are what we were trying to measure, so I call them relevant variance. Segments 2, 3, and 7 contain the various factors that we were not trying to measure, but that, nevertheless, consistently affect test scores or criterion measures. These are the main topic of the paper: irrelevant reliable variance.

The two segments numbered 8 are random error. How we implicitly define random error depends on how we measure reliability. I won't go into all of the aspects of random error.

Handout 2 names the variance components and lists some of the variables that influence them. The numbers of the various components are the same on handouts 1 and 2.

The part of variance that accounts for the validity coefficient is the football-shaped portion made up of segments 1 and 2, which is formed by the overlap of test and criterion measure. Segment 1 is the relevant part and segment 2 the irrelevant part of the variance common to the test and the criterion measure. This common variance is responsible for the correlation between the test and the criterion measure; that is, the validity coefficient.

Those characteristics of the examinees that are reflected in both the test and the criterion measure cause these variance components. Everything the test and criterion have in common that isn't job-related appears in segment 2. For example, a characteristic reflected in method variance on the test may also be reflected in a rater's perception of job performance. Having a large vocabulary may result in a higher test score and may lead to a higher criterion rating, while it may be "really" irrelevant to the quality of job performance.

The other part of irrelevant variance that concerns us appears in segment 3. This is test material that applicants respond to consistently, but that has nothing to do with job performance. This is the material that favors the test-wise, the fortunate, or the person who is in tune with the test writers. It allows applicants to get on the top of the hiring list for reasons irrelevant to the job.

Segment 4 contains any test factors related to job performance but not to the criterion measure. When we get a low validity

coefficient, we claim segment 4 is large. We say "The test is really a better measure of the criterion than our criterion measure is." And we frequently believe it, but nobody else does. Well, you can see it right here in the venn diagram. As an example, the test may inadvertently measure reading comprehension, which turns out to be important to job performance, but which we did not include in the criterion measure.

Perhaps the test measures the criterion better than the criterion measure does. But a good criterion measure likely measures the "real" criterion better than the test does. Segment 5 represents the part of job performance that is measured by the criterion measure and not by the test. You design your test to emphasize 1 and 4. You design your validity study to emphasize 1 and 5. If you do both well, 1 will be large and 4 and 5 will be small.

Segment 6 is the part of quality of job performance that is measured by neither the test nor the criterion. You can never measure how big this is. How well you get at the real criterion is determined only by judgement.

You could have a large overlap between test and criterion measure and still have a large segment 6: a large part of the criterion that is not measured. For example, you could identify 12 job elements in a job analysis and decide to measure only one of them. You could measure that one perfectly and still not measure much of job performance. Specifically, of all the things a secretary does, you could test for typing speed and validate against typing performance. A validity coefficient of 1.0 would not assure good prediction of the job performance described in the job analysis.

Segment 7 represents the unique part of the criterion measure, the part that is associated with neither the test nor the job performance. Segments 2 and 7 together constitute the most frequent flaw in validity studies; the failure of the criterion measure to represent the real criterion. This occurrence attenuates the validity coefficient. This attenuation can not be corrected for by the statistical correction for attenuation. That formula considers only the attenuation due to random error.

Segments 2 and 3 include the irrelevant test variance that we want to reduce. These are the variables that bias our test results. Be aware that when we reduce these components we lower reliability, because we reduce total variance without reducing random error. At the same time we increase validity, because the relevant variance is now a larger proportion of total variance.

The formulas at the bottom of handout 1 show this phenomenon. The reliability coefficient, r(xx), will increase if you add the variance of segments 1, 2, 3, or 4. The diagram illustrates the

same concept. If segments 1, 2, 3, and 4 are increased while random error stays the same, the test will have proportionately less error and will appear more reliable. Only random error adversely affects reliability coefficients.

We have been told that we should keep our reliability as high as possible. I'm telling you that is not necessarily so. When the reliability is the result of irrelevant variance it is of no use. It is worse than of no use. It makes our tests unfair. I would rather the non-relevant variance be error variance and lower the reliability coefficient, than to have variance that favors who knows whom. Whether the variance favors Shakespeare buffs, people who have taken introductory psychology, or truck drivers, if it is not related to job performance, it should not be in the test.

Selecting items using item analysis against total score can contribute to an unwanted reliability. If total score contains a good share of irrelevant variance, item analysis will identify the items consistent with the irrelevant variance.

The validity formula , r(xy), shows that validity is the common variance divided by the product of the square roots of the total variances. If the total variance of either the test or the criterion measure goes up, without an increase in the common elements of segments 1 or 2, the validity coefficient goes down. What's more, the validity coefficient will look better if you increase the shared irrelevant variance in segment 2. In the first two papers we were talking about increasing validity by decreasing component 3, irrelevant test variance.

You can also increase the validity coefficient by decreasing component 4: that is by removing relevant material from the test that is not included in the criterion measure.

What happens in a meta analysis of validity studies? It is likely that the variance common to a wide variety of settings is irrelevant variance of the kinds we have been talking about. This implies a caution concerning validity generalization. The validity coefficient being generalized may contain a large dose of shared irrelevant variance.

We must be judicious when we use validity coefficients to demonstrate that our tests are valid. We may be fooling ourselves consistently. We may have some blind spots or misconceptions that apply equally well to the test and to the criterion measure. Our tests include method variance, test-wiseness, and cultural bias, which increase the reliability of our tests at the expense of job-relatedness.

* * * * * * * * *

THE DESIGN AND APPLICATION

OF THE PROMOTABILITY INDEX

Elizabeth Mackall, Assistant Director
Public Sector Services
Personnel Decisions, Inc.

Introduction

Personnel Decisions in an I/O Psychology consulting firm based in Minneapolis. In our Pubic Sector Services Division we work with a large variety of organizations, ranging from large state jurisdictions such as the State of New York and the State of California, down to tiny mun' ipalities with populations under 5,000. Although we've worked with the full gamut of classifications from key executive, such as city manager, to custodian worker positions, our work in the area of selection and promotion tends to be predominantly with protective service classifications such as police, fire and corrections.

Typically, when we present at IPMAAC, we describe work we've been doing with large jurisdictions, such as the written simulations we've worked on for the States of California and New York. Today, I'd like to share some of the work we've been doing with small jurisdictions in Minnesota in the area of police promotions, specifically the Promotability Index we have developed for the Police Sergeant rank.

Small Police Jurisdictions and How We Work With Them

In our work with small police jurisdictions, we've learned to anticipate a number of common factors will be present, and will influence the kind of assistance we provide.

1. On the positive side, because the departments are small, and have a simplified structure, relationships within the department are fairly intimate -- everybody knows and works with everybody.

2. Also on the positive side, formal litigation or challenge to the promotional process is almost unheard of.

3. On the negative side, budget money available for the promotional process is quite limited, yet the stakes involved, from the perspectives of both the candidates and key department and city administrators, are just as high or perhaps even higher than in large jurisdictions (possibly stakes are higher because of fewer opportunities for promotions, and high visibility in the community).

4. Finally, on the positive side, there is a very high degree of commonality in job duties and requirements across jurisdictions. This mitigates to a large extent the dilemma posed by balancing the need for high quality promotional procedures with severe budget limitations. That is, after conducting a confirmatory job analysis and a workshop session with SME's to discuss contextual elements, such as community problems and issues and specific departmental organization, policies and procedures, almost without exception we find that the instruments and materials we have available need very little modification to make them consistent with the setting and job requirements in the specific jurisdiction. In the past few years we have developed a number of parallel or alternate forms of each exercise or testing instrument; thus we can offer the jurisdiction a cafeteria style menu, and assist them in selecting the procedures that best suit its particular needs and budget or administrative constraints.

The Promotability Index

We developed the Promotability Index late last summer when we were preparing Sergeant Promotional systems virtually simultaneously for three fairly small departments in Minnesota. each wanted some way of incorporating a performance appraisal into the testing matrix.

In our work with large jurisdictions up to this point, it seems if we so much as mention the word "performance appraisal" we are met with extreme hostility form at least one quarter - administration, the union, minority groups, etc. -- so we have never been able to introduce it other than as a criterion validation measure.

In the small departments we were working with last summer, however, it seemed perfectly reasonable and sensible that past performance be part of the promotional equation. Although this seemed reasonable to us as well, it was also clear that performance in general was not the issue of concern. Rather, the issues to be addressed would have to be performance on those aspects in the current job that would carry over and be critical determinants of success in the promotional position. To identify these, we met individually with representatives from the three different departments, reviewed the job analysis results for the Sergeant rank, and discussed to what degree and how each of the critical performance dimensions identified for the Sergeant rank could be observed at the officer level. We came up with five such performance dimensions. These are listed and defined in Handout A.

Following this, we developed a two step procedure for rating candidates on each of the performance dimensions. The first step involves assigning each candidate to one of 5 broad categories or levels of performance for that dimension, ranging from Very Good or Very Well at the high end to Very Poor or Very Poorly at the low end. (Show Handout B, column A.) Once all candidates have been placed in one of the five levels or categories, the second step for the rater involves rank ordering the candidates within each category (Show Handout B, Column B). The raters must complete both rating steps for all candidates on a particular dimension before proceeding to the next dimension.

Because the departments have been quite small and the working relationships among the various levels fairly intimate, in the jurisdictions that have thus far administered the Promotability Index for the Sergeant rank, virtually all supervisory and command staff from the rank of Sergeant through Chief have participated in the ratings. On a couple of occasions there have been as many as or almost as many raters as there have been candidates rated. With the ratings forms, however, each rater is given a sheet where he can list candidates whom he feels he is unable to rate on a particular dimension. Before the rating process is administered, raters are encouraged to rate a candidate on a dimension only if they feel fully confident that they are familiar with that person's performance on that particular dimension.

Prior to administering the Promotability Index, we train the raters in a group session. First we go through the Performance Dimensions, discuss the definitions and anchors for each, and have the raters brainstorm behavior examples that they believe fit a particular dimension. The purpose of the brainstorming is in part to flesh out the definitions and anchors for each dimension with descriptive examples. In part its to encourage a common perspective, so that all raters are attending to the same aspects of behavior when rating a particular behavior, as a behavior involves multiple components, each belonging to a different dimension, and sometimes the behavior involves multiple components, each belonging to a different dimension.

After the brainstorming process has been completed, we discuss typical errors in rating, such as halo, central tendency or leniency, and allowing personal preferences and prejudices to influence observations of behavior. At the conclusion of this discussion, the rating process begins. Each rater is instructed to work independently, and to complete all ratings for a particular dimension before proceeding to the next dimension.

Outcomes

Since we began offering the Promotability Index, it has been administered as a promotional testing device in five different

jurisdiction, and is planned for four more within the next month or two. Thus far, we have been fairly pleased with its results. It has been well received, not only by the department administrators (not surprisingly since they have been directly involved) and by the police and/or Civil Service commissions, but also but by the candidates themselves, several of whom have said it was fair even though they have been disappointed with the results.

Inter-rater reliability, however, although acceptable, has not been as high as we have initially hoped. We have hoped that the independent two-step rating process (categorization then ranking), combined with rater training would produce high inter-rater reliability, while at the same time minimizing halo. This hasn't been as much the case as we have initially hoped for. The overall inter-rater reliability coefficients for the four jurisdictions for which we have data have all been in the .67 to .76 range. On a dimension by dimension basis, the coefficients have raged from .63 to .90. Thus reliability of the Promotability Index is somewhat higher than for other performance appraisal systems we've encountered, but is still substantially lower than other testing devises we use, such as the behavioral oral and written job knowledge test.

On the other hand, when we correlate the results of the promotional potential with the results of other devices, we have found some stability across jurisdictions that has been encouraging. (show Handout 3).

The data shown on Handout 3 comes from four police jurisdictions in Minnesota who have used the Promotability Index in their Police Sergeant Promotional Systems.

In the handout, the uncorrected correlation coefficients of the Promotability Index with three other testing devices are shown. As can be seen from the handout, the strongest and most stable correlations are with the behavioral interview.

The uncorrected coefficients for the promotability index with the behavioral oral range from .43 to .70. Because the departments are quite small, the number of candidates involved are tiny, ranging from 7 to no more than 10, so the individual coefficients by themselves are not statistically significant. To get an idea of the strength of the correlation across all jurisdictions we merged the four data files together. At this point we had to adjust the obtained scores to compensate for difference in means and standard deviations among the four jurisdictions. We used an unadjusted linear transformation technique that has been in use since the 1920's. This technique preserves more of the true distributional properties of the data than does Z scoring, and is the same method we have used for adjusting scores for oral examinations in large jurisdictions such as San Francisco where

multiple panels are needed. When scores are adjusted to compensate for differences across jurisdictions, the correlation coefficient obtained between the Promotability Index and the Oral Interview is .57. This is quite respectable, and since it is based on 34 cases, is statistically significant at the .01 level. Admittedly this is still a very small number. However, given the relative stability of the separate coefficients for each of the four jurisdictions, we anticipate that as Promotability Index continues to be used, and cases are added, this level of relationship will tend to be maintained.

Indeed, in the fifth jurisdiction that has used the Promotability Index in combination with a behavioral oral supplied by us, the results appear to be quite consistent. Unfortunately, we do not have the raw score data from that jurisdiction, only the rank-order standings of the candidates on each of the testing devises used. In this jurisdiction, of the six candidates who have proceeded to the oral interview phase, there is a perfect correspondence between their rank-order standing on the Promotability Index, and their standing on the Oral Interview. That is, the candidate receiving the highest score on the Promotability, received the highest score on the Oral Interview; the second highest score on the Promotability was the second highest on the oral; and so on down the line.

The relationship between the Promotability Index and the behavioral oral is interesting for several reasons. The behavioral interview and the Promotability Index are designed to measure several of the same aspects of performance. Hence, it would be anticipated that they show a reasonably strong relationship. The two devices go about the process of measurement so differently that it is encouraging that the anticipated relationship does hold up.

1. in the Promotability Index, performance is rated by individuals who have worked closely with the candidate of a long period of time; by contrast, the panelists in the oral interview have not had previous contact with the candidate prior to the interview; and in the interview itself, the duration of contact is no more than 45 minutes.

2. in the Promotability Index, raters are specifically instructed to consider all past behavior relevant to a particular dimension, and not to focus solely on an exceptional or recent incident; by contrast, the behavioral interview specifically focuses on exceptional or recent incidents, by phrasing questions in terms of "the last time: or the "best", the "worst", the "most" and so on.

3. in the Promotability Index, the behavior rated is that observed directly by the raters; by contrast, in the behavioral interview, with the exception of the oral

communication dimension, the behavior rated is not observed directly by the raters but is reported orally by the candidate who is attempting to portray himself or herself in a favorable light.

The relationship between the Promotability Index and other testing devices is less clear because it is less stable. In general, the Promotability appears to have a somewhat negative relationship with the in-basket exercise -- in three of the jurisdictions for which we have data, the relationship is negative, but in the fourth, it is fairly strongly positive. We hypothesize that over time, as we add cases to the file, the relationship between the Promotability Index and the Oral will remain insignificant. That is, although there is some overlap in what is being measured, the Promotability Index is designed specifically to tap those aspects of performance that carry over from the Officer level to the Sergeant level, while the in-basket is designed to tap several abilities critical to the Sergeant rank that may not be needed or observed at the Officer level.

The correlation coefficients between the Promotability Index and the job knowledge test are also mixed. In two of the three jurisdictions for which we have data, they are fairly to strongly positive (.59 to .81). For the third jurisdiction, however, the coefficient is essentially zero. It should be noted that for the third jurisdiction, much of the job knowledge test was constructed in-house by the Police Chief and his command staff; it was not well received by the candidates, and numerous questions were appealed as trivial or overly technical. It is our hypothesis that over the long run, when the Promotability Index is used in conjunction with a well constructed and internally consistent job knowledge test, the coefficients will be significantly, but probably not very strongly, positive.

In summary, we've found the Promotability Index to be a useful and well received addition to the promotion systems we have implemented in small jurisdictions for Police Sergeant. We plan to continue monitoring its reliability in the hope of discovering whether there might be ways of improving inter-rater consistency. In the near future, variations of the device will be used in fire promotions in two different jurisdictions in Minnesota and a peer rating version of the process will be used in a moderately small sized jurisdiction.

* * * * * * * * * * *

SELECTING 911 CENTER TELEPHONE OPERATORS

WITH A MULTIPLE HURDLE EXAMINATION

Judith Trabert, Thomas Johnson, Sally Gale
City of Rochester, New York

Introduction

What do you do when:

Your city's Emergency Communication Center (911) operators have an attrition rate of 30% in the first year?

Operators hired for their clerical skills can't cope with callers who get hysterical or use foul language?

Candidates on the eligible list decline the job in droves when it is described during their interview with the appointing authority?

You modify your selection process. This paper describes the redesign of a multiple hurdle selection process in order to streamline the hiring of entry level 911 telephone operators-Telecommunicators. The paper presents the logic behind our redesign and some preliminary results.

Context

Rochester, New York is an upstate city with a metropolitan population of 900,000. Telecommunicators are civilian telephone operators who take information from callers in city and suburban areas on the 911 emergency services "hotline" of the Office of Emergency Communications (OEC) and pass it on to Dispatchers. They, in turn, direct police, firefighters and ambulance personnel to emergency situations, again in both city and county areas. Telecommunicators deal with long, boring periods of inactivity, work nights and weekends almost exclusively for their first several years of employment, and must be able to remain calm under pressure and when faced with abusive, hysterical or rambling callers.

In addition, Rochester's OEC is one of the most complicated 911 systems in the country in the number of agencies served, including police, fire and ambulance in the city and all surrounding towns. The computer-aided dispatch (CAD) system makes it necessary for Telecommunicators to learn between 300 and 400 computer codes (type of incident x type of response x agency), as well as a range of other skills, in an eight-week training period.

A selection process for the Telecommunicator title was first developed in 1983 and modified periodically in response to ongoing problems with recruitment, selection, and retention. At

the beginning of our project in 1987, the selection process consisted of: a written test which assessed clerical skills through subtests in directory usage, numerical sequence and transcribing information delivered orally; a typing test, and an oral performance test. The last was a job simulation in which candidates interacted with roleplayers via telephone to obtain information and complete response forms in emergency situations.

Problem and Proposed Solution

The redesign project was initiated when OEC management reported the following problems:

1. High turnover rate.
2. Many people put on the list by the old selection process were not employable but continued to block the list for weeks. Some turned the job down when its requirements were fully explained. Others were disqualified for medical problems or for an unsatisfactory police record.
3. A large percentage of new hires did not make it through the training.

Working with consultant Nancy Abrams, Ph.D., our staff concluded that these problems resulted from: misinformation about the job; screening too late in the selection process; screening inappropriately; and some skills, such as long term memory, not being tested at all. We decided that the process needed modification, not a complete overhaul, so we re-ordered and augmented the existing components. Our general solution had four parts:

1. To give more information about the job early in the process.
2. To screen earlier for bars to employment.
3. To change the minimum qualifications.
4. To supplement one component in order to better test memory.

Giving More Information

We had guessed that applicants often did not understand the stressful nature of the job, so we revised the examination announcement to include not only a description of typical work activities but also the following note:

> This job involves an unusual working environment which includes:
> *High stress of daily contact with life and death situations such as fires, murders, rapes and assaults in progress;
> *Close supervision and constant evaluation of work;

*Need to remain calm when speaking to people who are screaming, crying or hysterical;
*Need to remain polite with people who are angry, abusive or use foul language;
*Need to strictly follow rules and regulations.

Formerly, candidates were exposed to the mechanics of the job only at the post-list interview stage. Some candidates had been enthusiastic until they visited the job site; when they saw Telecommunicators on the job, that enthusiasm evaporated. Candidates now observe Telecommunicators at work for at least half a day, listening in on actual calls. At the end of the session, they are asked to sign a document indicating that they understand the nature of the job and are willing to work under its conditions. Both the description of working conditions and the observation sessions are intended to provide job preview information that encourages self-screening.

Earlier Screening

One change both provided job information earlier and screened more effectively. Previous examination announcements had stated that candidates must be available to work all shifts. But candidates weren't actually asked when they were available until the job interview, after the list had already been established. Because of the time it took to remove candidates from the list and contact those with lower ranking, delays in hiring occurred if candidates couldn't work all shifts. We suspected that many candidates didn't take that important job requirement seriously enough, so we moved the screen for "shift availability" up to the front of the selection process. In the first stage after application review, candidates complete a questionnaire about their ability to work rotating shifts, weekends and holidays, and other non-standard schedules. Applicants who answer any question in the negative don't proceed further in the selection process. Use of the questionnaire serves the double purpose of alerting candidates early on to the non-negotiable job requirement, and screening out unavailable applicants before they or we have invested much time or effort in the process.

In the earlier selection process, medical exams and police record checks were done after the list had been established, providing another way for ineligible candidates to block the list. We moved these components into the exam process itself, as another effort to reduce post list hiring delays. By securing candidates who were "appointment ready", we speeded up the post-list activities leading to employment.

Revised Minimum Qualifications

The third and most complex question was what population to recruit from and what to screen for in the minimum qualifica-

tions. In the past, applicants had been screened for clerical skills and experience in interviewing, explaining to, directing or informing the public, as evidenced by positions such as complaint clerk, receptionist or salesperson. These provide opportunities for face-to-face encounters, in which a significant portion of the information exchange takes place through gestures or body language. The Telecommunicator position, however, requires effective interaction with an unseen caller. So we revised the minimum qualifications to target candidates who had worked in a stressful environment which also included indirect communication or emergency situations.

The new requirement asks for six months of paid or volunteer experience interacting with the public using telephone, two-way radio, or other means of indirect communication in an emergency or other setting in which speed of response is critical. Such experience might be gained as a public safety telephone operator or dispatcher, a hospital or medical answering service operator or taxi dispatcher. Alternatively, candidates can have six months of experience with the public face to face in an emergency setting in positions such as emergency medical technician (EMT), firefighter, or ambulance technician.

Early on in the conceptualization of the Telecommunicator position, it had been seen as primarily a clerical job. As the position and its selection process evolved, the clerical emphasis decreased. Over time we had discovered that a strong clerical bias screened out applicants with emergency service experience, but that the absence of any typing requirement produced candidates who couldn't learn interaction with the computer terminal fast enough. The new exam includes an assessment of typing skill, but de-emphasizes its level and source. The Keyboard Familiarity subtest is designed to evaluate minimum skill level on a typewriter or computer-style keyboard. Speed is not a primary consideration and the required skill level is not spelled out in the announcement. To encourage non-professional typists to participate, the announcement states that "typing technique is NOT important" and that "hunt and peck" typing is an acceptable style. The clerical bias of previous minimum qualifications and exams has been removed in order to focus on a more appropriate candidate population. Those candidates who otherwise meet or exceed the job qualifications need not be discouraged from applying simply because they are not proficient typists.

The rating of training and experience (mini-T&E) is a completely new test component which uses and builds on the minimum qualifications. Candidates are given ranking points for combinations of experience such as experience with indirect communication with the public in an emergency setting or indirect communication and separate experience working in an emergency setting. In addition, candidates are given additional points based on their fluency level in languages other than English.

The last part of the questionnaire gives candidates credit for "computer familiarity" - experience using a computer at work (paid or volunteer), school or home. This credits candidates who are comfortable with computers, since some previous appointees had suffered from computer phobia. Although the training and experience section isn't a pivotal component of the examination, it rewards candidates who have developed job-related skills which could enhance their performance as Telecommunicators.

Study Guide

Finally, we were concerned with assessing candidates' recall abilities because of the large matrix of codes which Telecommunicators must learn. So, two weeks before the oral performance and typing subtest, we sent candidates a study guide. The first part of the guide included practice typing materials. A second part of the study guide prepared candidates for the oral performance subtest. They were asked to familiarize themselves with the procedures used in responding to calls and the guidelines used for different kinds of incidents. In addition, they were asked to memorize a list of incident codes that they would use in the actual exam. In the exam, candidates were asked to respond to three simulated calls by interacting with roleplayers over the phone and by filling out a report form similar to the one used on the job. This section of the exam tested the candidates' abilities to elicit information, to reply in a professional manner in a stressful situation, and to accurately remember and write information they receive over the telephone.

The oral performance subtest represented the final phase in this multiple hurdle exam. In its amended and expanded form, the redesigned selection process consisted of an evaluation of availability, a mini T&E, a police records check, an onsite observation session, a keyboard familiarity subtest, an oral performance test, and a medical exam, in that order.

Results and Plans

Our results are mixed. We wish we could say that we had eliminated all of our problems, but we can't. We seem to have uncovered some new ones, in that the new exam has high adverse impact. The new minimum qualifications tend to favor suburban volunteer firefighters and ambulance personnel; most of Rochester's minority population lives in the city. In general, the more stringent minimum qualifications may have discouraged applicants, since the agency's recruitment difficulties have not abated.

In spite of these concerns, our redesign did produce faster appointments, less list blockage, and higher percentage of candidates who completed training, in a job in which constant stress and high turnover are endemic.

Under consideration as future steps in the redevelopment process are: a new job analysis to better reflect the demands of the CAD system; a keyboard familiarity test administered on a computer keyboard with information delivered aurally to better simulate job conditions; and the consideration of personality factors which might distinguish candidates suited for the Telecommunicator job.

* * * * * * * * * * *

A "MAILABLE COPY" TYPING TEST

Nelson Adrian
Robyn Wachtel
Steve Magel
T. R. Lin

Los Angeles Unified School District

This paper reviews the development of a specialized work sample typing test for secretarial candidates. The test goes beyond the traditional "straight copy" typing test that assesses a candidate's ability to type with speed and accuracy. This mailable copy typing test also measures candidates' ability to set up and type a letter suitable for mailing from an unformatted handwritten copy. Successful candidates must be able to type a letter quickly and accurately, proofread and correct errors, correct typing mistakes, set proper margins and salutation, and close letters in the same manner as they would be required to do on the job.

Test Development

Background

In 1984, the classification of Secretary was divided into two separate classes - Stenographic Secretary and non-Stenographic Secretary. After the division, non-Stenographic Secretarial candidates were not required to take a stenographic test. Unfortunately, some administrators found that individuals hired from the non-Stenographic Secretary list were often unable to perform basic secretarial duties such as setting up and typing business letters, proofreading, etc.

Taking this into consideration, we judged it necessary to develop a job related, work sample performance test that would assess the ability to prepare and type business correspondence; to proofread accurately; and to follow directions.

Job Analysis

A series of job analysis interviews were conducted in order to determine specifically what supervisors expected of a secretary in terms of typing ability. It was determined from these job analysis interviews that secretaries are frequently asked to type business correspondence from a hand written draft not set up in letter format. Further administrators often expect their secretary to proofread letters, and independently correct any punctuation, capitalization, or spelling errors. Secretaries are expected to make these corrections without assistance.

The Concept of Mailability

On the surface, the concept of mailability sounds as though it would be simple to define and measure. However, there is no concrete definition of mailable; rather, it is more a matter of judgement. In an attempt to define this concept more precisely, letters were typed with various errors. Judges were asked to rate each of the letters in terms of their acceptability as "mailable copy". The results were found to depend on instructions regarding whether the letters were described as test material or not. In this case, the majority found letters with 6 errors to be, at least, barely passing. Considering these results and factors such as test taking anxiety, unfamiliarity with the typewriters used during the exam, and machine peculiarities, it is unrealistic to expect typists to produce three perfect letters under examination conditions. Thus, the concept of mailability was defined as letters that have errors which can be corrected without causing the finished product to appear sloppy.

Performance Test Description

The actual performance test is comprised of three hand written letters which must be typed within a 30 minute time limit. The candidates are also given a five minute practice session with a sample handwritten letter. The handwriting of the letters is intended to be neat and clear. Three different forms of the test have been developed. While each letter varies in content, each form has comparable letters which are approximately equal in terms of number of sentences (5 to 7), number of words (140 to 159), number of strokes (887 to 945) and FOG Index difficulty (7.75 to 10.42). Each letter contains three "planted" punctuation, capitalization, or spelling errors which candidates are to correct or points will be deducted for typing the mistake.

It should be noted that while this mailable copy typing test does take a bit longer to administer than a standard typing test (about one per hour), the length and time limit still allows for numerous testing sessions to be scheduled in one day.

An administration manual has been developed to accompany the test. This consists of : Instructions for Candidates, Instructions for Proctors (those administering the test), Instructions

for Raters (those scoring the test), Examples of scored letters with errors, A Margin Guide, and An Error Guide.

The purpose for this manual is to insure that the instructions to candidates, administration procedures, and scoring procedures are standardized across administrations. The Error Guide consists of a comprehensive list of, essentially every conceivable typing error, a typed example of the error for further clarification, and the number of points which should be deducted for each error.

Scoring

Scoring System Development

After reviewing several references and our definition of mailable, we decided that the number of points deducted for an error committed by a candidate should be related to the proportion of time required to retype or correct the mistake, enabling the finished letter to be mailable. Consequently, we followed these guides which require one point to be deducted for each minor error (i.e., a missing letter or an extra space), and three to four points for each major error (i.e. omission of a word or typing a word twice). Some errors such as skipping a sentence may not be correctable, but are considered as only one mistake. Consequently, only four points are subtracted.

Rater Training

An additional step was taken to insure the proper scoring of the typing products of candidates. A rater training session was developed. Briefly, this session consists of a discussion of the purpose of the training followed by a complete review of the error guide (step by step, one error at a time). This includes soliciting and answering questions until each point is understood. Clarification of errors requiring judgement, such as scoring erasures, are discussed in detail.

Pass Point

The pass point for this type of mailable copy test may be modified to suit one's business needs. However, based on the ratings obtained from LAUSD's administrators and our demand for secretaries, the pass point was set at 21 points off for the total of three letters. The pass rate for our candidate population has been about 60% at this pass point.

Evaluation of Reliability and Validity

As is apparent from this discussion and the test development process, the primary validity evidence for this performance (work sample) test is content validity. Candidates are asked to type letters similar to letters they might type on the job. To be successful they must demonstrate basic skills relating to following directions, proofreading, setting up and typing

correspondence at a minimally acceptable level as determined by supervisors of these secretaries.

One important point concerns the complexity of this exam's scoring procedure. In order to determine if raters were having difficulty scoring this test as opposed to the traditional typing test, we took one hundred typed letters from a standard typing test and one hundred letters from the mailable copy typing test and carefully scored them a second time. An inter-rater or score-rescore reliability of .86 for the standard typing test and a reliability of .90 was found for the mailable copy test. This greater reliability coefficient for the Mailable Copy Typing Test may be a result of the training session, the Error Guide, and /or the fact that the mailable copy test may have been scored more carefully because of the attention and novelty of the exam. In any event the raters do not appear to be having difficulty with the scoring.

Finally, it should be mentioned that we would like to collect performance scores in the future to further validate this test, as we believe it warrants. However, time constraints have made this impossible at this point. In the future we also plan to further evaluate our scoring system by examining how many errors of each type are made and how many points are deducted for these. Also, since many Secretaries have access to computers and word processing software, we plan to consider adapting this exam or this format for use in a word processor context.

* * * * * * * * * * *

DIRECT VS. INDIRECT ASSESSMENT OF WRITING SKILLS:

A LOOK AT SOME OF THE LITERATURE

Michael J. Dollard
Principal Personnel Examiner
New York State Department of Civil Service

The paper looks first at a nationwide survey conducted by the New York State Department of Civil Service. It consisted of a multi-page survey instrument distributed to 70 public and third-sector county, state, and quasi-public organizations from across the country. Twenty percent of the organizations surveyed do not test writing skills at all. Of the 80% which do, there is a great variety of practice. Almost all of them use some form of indirect assessment (primarily some form of machine-scored multiple-choice test) and fully two thirds of them use direct writing assessment (i.e., writing samples) as well. The job

groups for which writing assessment is most frequently used are clerical operatives (e.g., clerks, typists, etc.), clerical supervisors, secretaries, administrative staff and mangers.

Of those organizations using direct assessment, about half use a single writing sample, but half use multiple samples, usually two or three. A variety of rating methods are used, with about 20% using "holistic scoring" and the remainder using some type of "point-factor" rating. Common rating criteria are quantity and quality of ideas, clarity, grammar and usage, appropriateness to purpose, organization, and clarity. Of those organizations using indirect assessment with multiple-choice items, nearly all use some form of "grammar", "English usage" and "vocabulary" items, as well as some type of editing of sentences or paragraphs. In the evaluation of writing skills for "professional" (i.e., college educated) job types, the most common objective test types are the construction shift, sentence completion, and scrambled paragraph items. Candidate populations vary widely in size, with direct assessment methods being used on populations from two or three up to 12,000! Indirect assessment is used with an even broader range, up to "tens of thousands" in some cases.

A literature search was conducted but found little published material, and even less unpublished material. Peter Cooper did a literature search a few years ago for the Graduate Record Examinations Board. His conclusion summarizes what we also found: The literature indicates that writing samples are often considered more valid than multiple-choice tests as measures of writing ability. Certainly they are favored by English teachers. But although writing samples may sample a wider range of composition skills, the variance in such scores can reflect such irrelevant factors as speed and fluency under time pressure or even penmanship. Also writing sample scores are typically far less reliable than multiple-choice test scores. When writing sample scores are make more reliable through multiple assessments, or when statistical corrections for unreliability are applied, performance on multiple-choice measures, though, tend to overpredict the performance of minority candidates on writing samples. It is not certain whether multiple-choice tests have essentially the same predictive validates for candidates in different disciplines, where writing requirements may vary. Still, at all levels of education and ability, there appears to be a close relationship between performance on writing samples and multiple-choice test used to evaluate writing skills.

The Godshalk Study - 1966

In 1965/66 a team from ET, headed by Ferd Godshalk, undertook a comprehensive study of writing assessment for the College Entrance Examination Board (CEEB). This study involved the use of five different experimental writing samples, six objective

test types, two interlinear exercises and data obtained from two PSAT (Preliminary Scholastic Aptitude Test) essays administered under field rather that experimental conditions.

The criterion in the study consisted of specially designed writing samples covering five topics, each of which was rated by five carefully selected and trained raters. Two of the writing samples were somewhat lengthy (40 minute) exercises requiring analysis and planning, and some decision regarding interpretation, point of view, or a judgment that was to be stated or defended. The other three writing samples were much shorter (20 minute) exercise designed to elicit immediate response. The subject matter of the exercises was devised so as to stimulate different types of writing: descriptive, narrative, expository and argumentative.

A significant finding is the high subject by topic interaction, confirming that subjects do vary by topic in their writing abilities and suggesting that in any direct writing assessment, a variety of topics/writing samples must be provided to achieve even moderate reliabilities. A further implication of the moderate observed reliabilities is the cap which it creates for any demonstrated validity in the objective test predictors.

Eight predictors were used in the study: two interlinear exercises and six classes of multiple-choice questions:

paragraph organization
usage items
sentence correction items
paragraph completion items
error recognition items
construction shift items

All of the objective type tests were at least moderate predictors of the combined writing sample score as a criterion. Most inter-correlations among the objective test types are moderate, with Usage and Sentence Completion being the most highly inter-correlated with and intercorrelation of .775. The correlation of the sets of predictors range from .717 to .748, certainly respectable, and much higher that previously reported validities for writing tests.

The correlations between the two inter-liner exercises and the writing sample criterion were .651 and .597, in the same general range as the objective test types. In general, validities increase slightly when an inter-liner exercise is substituted for an objective test type other than Usage of Sentence Completion, and decreases slightly when substituted for one of these latter objective test types.

The field trial writing samples (i.e., PSAT essays administered and rated under field conditions), when added to the objective test type combinations, improved validity in proportion to the number of independent ratings they received, by even with four independent ratings they improved validity coefficients by only about .04.

In sum, the Godshalk team reached four conclusions from this study:

1) The reliability of writing samples is primarily a function of the number of different writing samples and the number of independent ratings included.

2) When objective test types specifically designed to measure writing skills are evaluated against a reliable criterion, they prove to be highly valid.

3) The most efficient predictor of a reliable direct measure of writing ability is one which includes a writing sample or inter-linear exercise in combination with objective test questions.

4) In the light of the small increase in validity provided by the addition of a writing sample or inter-linear exercise, it is doubtful that their addition can justify the large increase in administrative and rating costs which they entail.

The Breland Study -- 1987

In 1986/87 and ETS/CEEB team headed by Hunter Breland took another look at the assessment of writing skills. Initially intending to replicate the study done 20 years by the ETS/CEEB team headed by Godshalk, the Breland study ultimately went somewhat beyond the scope of the earlier study.

The Breland study used six writing samples by each examinee, two each in what are described as the narrative mode, the expository mode, and the persuasive mode. Each writing sample was rated holisticly by three independent raters, yielding 18 ratings per examinee. These ratings were combined to produce composite scores for each of the six topics and for all six topics taken together.

Although the Breland team had a greater variety of technology at their disposal that did the earlier team, and although they performed a greater variety of analysis that did the earlier team, their results largely replicate those of Godshalk, et al.

As the Godshalk team had concluded, the reliability of writing samples is directly related to the number of samples and

the number of ratings. The Breland team estimates the reliability of a single writing sample read once to be in the range of .36 - .46; read twice to be in the range of .47 - .57; and read thrice achieved reliabilities in the range commonly achieved by the multiple-choice tests such as the TSWE and ECT (i.e, .85-.92). The data would suggest that multiple-choice tests of writing skills are roughly equivalent in validity to single samples read twice or, preferable, thrice.

The Breland team, like the Godshalk team before them, performed a number of analyses to estimate the effect of using a combination of direct and indirect assessment methods. Breland's findings confirm preferably read two or three times, does improve the validity of assessment. The increment of improvement found by Breland is somewhat greater than that found by Godshalk.

* * * * * * * * * * *

THE DEVELOPMENT AND USE OF VIDEOTAPED WORK INCIDENT SIMULATIONS IN POLICE AND FIRE ASSESSMENT CENTERS

Betty M. Marshal and Jacqueline Page
Fairfax County, Virginia Office of Personnel

Videotaped work simulations based on actual job incidents were developed for four assessment centers: Fire Captain, Fire Battalion Chief, Police Lieutenant, and Police Sergeant. In the development of each work simulation, panels of subject matter experts identified realistic job situations which would require the application of knowledges and skills that had been identified through job analysis. Test development specialists, in conjunction with police and fire subject matter experts, wrote scripts for video vignettes to portray these incidents.

The four work simulations varied in content, length, and format depending on the type of assessment center exercise developed, the number of candidates to be assessed, and the purpose of the individual exercise in the total selection procedure.

All simulations were videotaped on location in various Fairfax County settings, such as restaurants or townhouse developments, using subject matter experts and other amateur volunteers as actors.

Since Fairfax County has an internal Cable TV Programming division, video equipment and expertise were available at no out-of-pocket cost to our office. Trained staff of the Police and Fire Departments did most of the actual filming and editing using department-owned equipment or borrowed from the Cable TV division.

The use of the video format was first proposed to eliminate the problems of inconsistency and actor fatigue that are often experienced with role play exercises, particularly with a large candidate population, and to allow development of exercises more closely related to the job. The format was then applied to the incident simulation exercise for these same reasons and to reduce administration time and effort. The use of the video format resulted in:

- -increased job relatedness
- -greater standardization of exercises, instructions and administration
- -increased candidate acceptance
- -the opportunity to present more complex job situations that real time simulation or paper and pencil tests would allow

The major weaknesses of the video format were:

- -the increased time and technical requirements for exercise development, and
- -the lack of direct feedback in response to candidate actions (as compared to role-play exercises)

Following is a brief description of each of the simulations developed including the purpose of the exercise, the dimensions examined and the number of candidates assessed. This is followed by a discussion and administration phases.

Fire Captain Interaction Exercise

Setting: Fire Station

The simulation is a videotaped series of five (5) encounters between the off-camera station caption and a number of subordinate employees. The candidate assumes the role of the station captain.

Purpose: To test the candidate's skill in problem-solving and supervision.

Response Format: Written response

Candidates responded in writing to each of the five scenes, identifying the issues involved, describing any immediate action they would take and any follow-up action required. After responding to the five individual scenes, candidates identified major issues

and concerns and long term actions needed to resolve them.

This exercise was administered to 15 candidates in a single sitting.

Dimensions Examined:
-Analysis
-Relationship with People
-Supervision
-Commitment to Management Role
-Communication
-Behavior Under Stress

Fire Battalion Chief Interaction Exercise

Setting: Fire Station

This simulation is a series of scenes from an unanticipated meeting between a Battalion Chief and a couple that has dropped by the station to follow up on a complaint filed with the station captain three weeks prior. The candidate assumes the role of the Battalion Chief.

Purpose: To test the candidate's problem solving and interpersonal skills

Response Format: Written Response

Response required identification of issues, immediate and follow-up actions to resolve issues identified, and overall response to the problem.

This exercise was administered to 17 candidates in three small groups.

Dimensions Examined:
-Analysis
-Relationship with People
-Commitment to Management Role
-Management
-Communication
-Behavior Under Stress

Police Second Lieutenant Incident Management Exercise

Setting: Patrol

A Second Lieutenant on patrol in a Police vehicle

responds to the scene of a burglary and rape at a townhouse development.

Purpose: To test the candidate's skill in managing the incident supervising assigned squad members, and making a clear, concise oral incident report.

Response Format: Written response, oral presentation

Candidates were given 10 minutes prior to viewing the tape to review written background materials including and overview of the district, a squad line-up with backgrounds of squad members, a list of patrol areas with criminal activity by area, and an aerial map of the district. Candidates completed a written log of all actions taken, orders given and resources requested to handle the incident. Candidates then gave an oral debriefing report to two assessors as the duty captain and the public information officer.

This exercise was administered to 35 candidates individually so that the oral presentation could immediately follow the videotape.

Dimensions Examined:
-Application of Job Knowledge
-Decisiveness
-Interpersonal Relations
-Judgement
-Leadership
-Management Control
-Written and Oral Communication
-Planning and Organizing
-Behavior Under Stress

Police Sergeant Incident Management Exercise

Setting: Patrol

A Police Sergeant on patrol responds to the scene, where an armed robbery has just occurred.

Purpose: To test the candidate's skill in handling a basic incident on regular patrol.

Response Format: Oral Presentation
Candidates viewed the videotape of the incident then had 10 minutes to prepare a detailed 5-minute oral presentation of all actions taken and orders given to others to handle the incident.

This exercise was administered to 100 candidates individually so that the presentation could immediately follow the videotape. Presentations were audiotaped for later review by the assessors.

Dimensions Examined:

- -Analysis/Judgement
- -Decisiveness
- -Leadership
- -Oral Communication
- -Planning and Organizing
- -Behavior Under Stress

Results

In general, candidate response to the videotape exercises was positive. In feedback sessions, candidates in general expressed the opinion that the simulations were more representative of actual work situations than paper and pencil exercises. The built-in standardization of the video format eliminated complaints concerning mistiming and other administration errors.

Concerns raised in the development of these exercise can be grouped into the following broad categories:

- -Exercise content
- -Exercise development
- -Administration
- -Candidate training and orientation

Exercise content issues included exercise format, length and complexity, and the level of attention to detail the exercise required. We found that candidates watching a videotape are much more attuned to fine details and incidental background details than expected. This required that particular care be taken during the development phase to minimize or otherwise account for inconsistent background details.

The Exercise Development phase included script development, selection of actors, filming and editing, and security concerns. This phase was by far the most time-consuming and was the phase where most problems occurred. While development of the concept and general content of most exercises was fairly easy for test development staff and SME's, actual script development had to be extremely detailed.

Even though we had technical assistance from persons trained in filming and editing, the level of their experience was less than expert. We spent a lot of extra time learning new and easier techniques as we went along, and probably spent far more time on editing than a professional might have required. Hopefully, this acquired knowledge will carry over into future projects

Many of the roles in the simulations were played by SME's involved in the development process or by other volunteers. Since these were usually people with some association with our Police or Fire departments, security was a major concern. We also used in-house staff, both uniformed and civilian, as filming and editing technicians, opening another possible source of security leaks. Every effort was made to keep the number of people involved to a minimum.

Use of current employees as actors led to some unforeseen misinterpretations by candidates. This was usually the result of a candidate's knowing an actor in his real-life function and assuming that he served this same function in the simulation.

Administration was fairly easy when instructions and timing were built in to the videotape. Again, this required some special attention during the development phase.

Candidate training and orientation is an area where we feel more effort should be placed in future work simulation projects. Though feedback was generally positive, candidates were sometimes unclear as to the expectations of the assessors when making their responses.

Conclusion

Experience gained have raised issues to be considered in future assessment centers. These include training and preparation, timing of viewing and preparation, level of detail of visual presentation, and number of repetitions of the simulations exercise needed for clarity to candidates.

* * * * * * * * * * *

EXAMINATION OF EXISTING DATA TO PREDICT JOB PERFORMANCE FOR PERSONS WITH MENTAL RETARDATION

James S. Russell
The University of Oregon and Lewis and Clark College
and
Jon R. Lucke and Nancy Brawner-Jones
The University of Oregon

Abstract

Existing data were analyzed to determine job validities for cognitive, psychomotor, and social predictors for persons with mental retardation. In addition, job validity data from studies with private and public sector employees were analyzed to determine which characteristics of jobs reduced the validities of cognitive ability scores. Results indicated that cognitive, psychomotor, and social aptitude scores were highly correlated with various measures of job success in a variety of job settings. Results also identified twelve characteristics of jobs that caused cognitive validities to vary.

Introduction

The study was undertaken to summarize previous research on the job validities of cognitive, psychomotor, and social predictors for persons with mental retardation by using recent statistical techniques to cumulate data (Glass, 1982; and Hunter, Schmidt, & Jackson, 1982). Previous research on job validity studies had concluded that traditional assessment was of value in classifying individuals, but gave little guidance for persons who were responsible for training them (Cobb, 1972; Halpern, Lehmann, Irvin, & Heiry, 1982).

Another purpose of the study was to provide better guidance for social delivery systems working to place people with mental retardation by identifying a clear set of guidelines as to what characteristics of the job would minimize job support and training. The recent availability of detailed job dimensions data from the Position Analysis Questionnaire (PAQ; Mecham, McCormick, & Jeanneret, 1977) and other job validity data made it possible to analyze which job characteristics increased or decreased the validity of cognitive aptitude scores. This research was designed to expand on previous work which has established that jobs which require minimal decision making and processing of information decrease the job validity of cognitive aptitude scores (Gutenberg, Arvey, Osborn, & Jeanneret, 1983).

Method

A meta-analysis was conducted according to the procedures outlined in Glass, McGaw, & Smith, (1981), and Hunter, et al., (1982). A search through the library and personal contacts was made of published and unpublished research literature. A code book was established for coding the studies, and reliability statistics were established according to guidelines in Glass, et al., (1981) and Jackson (1980). A list of the studies that were used is available from the senior author.

The PAQ data were obtained from PAQ Services, Inc., and were merged with data provided from the U.S. Employment Service job validity studies based on the Dictionary of Occupational Titles and the General Aptitude Test Battery (DOT/GATB; U.S.

Department of Labor, 1970). The PAQ job data was matched with the DOT/GATB data for 438 studies on cognitive predictors and 436 studies with psychomotor predictors. Each study was weighed equally, with the validity correlations converted to Z values, and the validities corrected for restriction in range (Gutenberg, et al, 1983).

Results

The results of the study are separated into two sections. The first section is the summary of results from the meta-analysis of persons with mental retardation. The results of the cognitive studies show that global cognitive predictors have positive correlations with criteria that include supervisor ratings, employment status, wages, and job output in work settings that include sheltered workshops, supported employment, and competitive employment. The average correlation, corrected for restriction in range, is p = .47 (N = 3,472, K = 47 studies). The lower 90 percent confidence interval is p = .29, indicating that the correlation is significant. The social tests, such as the Vineland Social Maturity Scale, had an almost equally strong correlation (p = .45, N = 658, K = 5 studies). The psychomotor tests had an average correlation of .49 (N= 1,289, K =19 studies) without correction for restriction in range, which could not be calculated because data were not available. Sixteen of the nineteen pyschomotor studies were conducted in sheltered workshops on criteria of work samples, quality and supervisor rating.

The results indicated that the psychomotor scales have higher validities than either the cognitive or social scales and that cognitive and social scales appear to be equally effective at predicting job success. This does not imply that they are interchangeable; however, research in the mental retardation literature suggests they are complementary (Menchetti, Rusch, Owens, 1983).

The results for the analysis of the PAQ data are listed in Table 1. Table 1 describes the results of the correlations between individual job dimension ratings and job validities. The results indicate that there are 16 job dimension ratings out of 45 individual and overall job dimensions where the validities vary significantly according to the job dimension rating. Twelve of the cognitive correlations are positive, indicating that the validity increases as ratings on the job dimension increase, while four of the job dimension correlations are negative, indicating that the cognitive validities decrease as ratings on the job dimension increase. The pattern sign for the psychomotor scales is exactly the opposite; positive or negative correlations for cognitive scales are complemented with negative or positive correlations respectively for the psychomotor scales.

Discussion

The results from the meta-analysis indicate that cognitive, social and psychomotor test scores predict various measures of job performance in a wide variety of work settings. The results from the PAQ job dimensions indicate that cognitive requirements diminish when jobs are structured or require general body movements. The research should provide assistance to counselors working with persons with mental retardation by giving the counselors assurance that various assessment instruments can assist in the performance of people. The results also describe job characteristics that may be included in jobs to increase the likelihood of job success for persons with mental retardation. The results can be combined with research on utility theory to predict the economic impact for an employer who hires persons with mental retardation.

References

Cobb, H.V. (1972). The forecast of fulfillment: A review of predictive assessment of the adult retarded. New York: Teacher's College Press, Columbia University.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage Publishing.

Gutenberg R. L., Arvet R.D., Osborn, H.G., Jeanneret, P.R. (1983). Moderating effects of information processing/decision making on test validities. Journal of Applied Psychology, 68, 602-508.

Halpern, A.S., Lehmann, J.P., Irvin, L.K., Heiry, T. I. (1982). Contemporary assessment for mentally retarded adolescents and adults. Baltimore: University Park Press.

Hunter, J.E., Schmidt, F.L., Jackson, G.B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage Publishing.

Jackson, G.B. (1980). "Methods for integrative review," Review of Educational Research, 50, 438-460.

Mecham, R.C., McCormick, E.J., & Jeanneret, P.R. (1977). Technical manual for the Position Analysis Questionnaire. Logan, UT: PAQ Services, Inc.

Menchetti, B.H., Rusch, R.R., Owens, D.M. (1983). Vocational training, in Matson, J.L. & Breuning, S.E. (Eds.), Assessing the mentally retarded, (pp 247-284). New York: Grune & Stratton.

U.S. Department of Labor (1970). Manual for the General Aptitude Test Battery. Washington, D.C.: U.S. Government Printing Office.

Table 1

CORRELATIONS BETWEEN PAQ JOB DIMENSIONS AND GATB COGNITIVE AND PSYCHOMOTOR JOB VALIDITIES[1]

Division	Job Dimension		Cognitive	Psychomotor
			(n=438)	(n=436)
I.	2.	Using various sources of information	.26	-.26
II.	7.	Making decisions	.24	-.24
	8.	Processing information	.18	-.18
III.	10.	Performing activities requiring general body movements	-.13[4]	.06[2]
	12.	Performing skilled/technical	.17	-.20
IV.	17.	Communicating judgements/related information	.22	-.20
	20.	Exchanging job-related information	.12[4]	-.09[3]
V.	23.	Engaging in personally demanding situations	.17	-.17
VI.	26.	Working in businesslike situations	.23	-.23
	29.	Working on a regular vs irregular schedule	.17	-.12[4]
	30.	Working under job-demanding circumstances	.21	-.25
	31.	Performing structured or unstructured work	-.19	+.16
VII.	33.	Having decision, communication and general responsibilities	.24	-.23
VIII.	35.	Performing clerical and related activities	.18	-.24

IX.	39.	Performing routine activities/ repetitive work	-.15	.22
X.	45.	Unnamed	-.18	.11[4]

[1]All validities are significant at $p < .001$ unless noted
[2]$p > .10$, n.s.
[3]$p = .04$
[4]$p = .01$

* * * * * * * * * * *

EMPLOYMENT OF THE DISABLED:

ACCOMMODATING PEOPLE IN THE WORKPLACE

James Breene, Senior Support Center Representative
IBM Corporation
Marietta, Georgia

In the United States today, there are 36 million Americans who are identified as disabled, according to the U.S. Census Bureau report. Of this number, 48.2% or 17.2 million are in the 16-64 working age population. And...there are 500,000 people being added to the total disabled count each year.

The cost of disability support is staggering. We are looking at $119.6 BILLION through a variety of federal, state, and private support payment structures. In comparison, in the same period of time, $3 BILLION was spent on rehabilitation. That's 2.5% directed toward creating independence, self-respect and a self-support structure for our people.

What is a disability? According to the Federal Rehabilitation Act of 1973, "a disabled person is a person who has a physical or mental impairment which substantially limits one or more of his or her life activities". Major life activities are: self-care, socialization, education, transportation, housing and more particular for our purposes, employment. Studies have shown that almost 70% of disabled men and 81% of disabled women are not employed. The numbers are even worse for disabled minority Americans.

A handicap is really an interaction between a disability and an environment. The person has the disability. The person works in an environmental setting. This is to say, if an environment is modified so as to be non-handicapping, the person is really no longer handicapped. He or she may still be disabled, but no longer handicapped. There is not a barrier, not an impediment, not something in the environment that keeps that person from functioning.

The environment variables fall into several categories:

Attitudes
- The disabled persons'
- The non-disabled persons'
- The organizations'

Accessibility
- Access to a company as a place to work
- Barrier free access to one's workplace/work station

Accommodations
- The willingness and creativity displayed in the way we do things, the way we arrange things, the way we equip qualified disabled individuals to do their jobs despite limitations
- The use of available technology to provide a disabled person with the ability to function in a competitively employed capacity

Can a person with a disability perform up to the expected work standards of a business? Let me share some facts from the E.I. Dupont de Nemours Company. They conducted a study of 1958, updated it in 1973, and re-validated the study in 1981. The study in 1981 involved 2,745 disabled Dupont employees. The following is from the Dupont study,

	Disabled Employees	Non-disabled Employees
Performance	92% Avg to above avg	91% Avg to above avg
Safety Record	96% Avg to above avg	92% Avg to above avg
Attendance	85% Avg to above avg	91% Avg to above avg
Turnover	Considerable less than non-disabled employees	

A similar study was done by the International Center for the Disabled in cooperation with the National Council on the Handicapped and the President's Committee on Employment of the Handicapped. The survey results were published by Louis Harris and Associates, Inc. in March 1987 in the document "The ICD Survey II: Employing Disabled Americans. In Chapter 5 on page 45 under the heading: Managers Rate the Job Performance of Disabled Employees, the following is quoted: "Overwhelming majorities of top managers, EEO officers, department heads/line managers, and small business managers give disabled employees a good or excellent rating on their overall performance. Only one in twenty managers say that disabled employees' job performance is only fair, and virtually no one says that they do their jobs poorly".

Should the cost of accommodations be a factor in employing a person with disability? In 1983, the U.S. Department of Labor-Employment Standard Administration commissioned a survey that was conducted by Berkley Associates. The survey covered approximately 20,000 disabled employees of those firms. This survey found that 51.1% of the accommodations had no associated cost. For 18.5% of the accommodations, the cost was $1-99. That's almost 70% cost less than $100.00. The other breakouts were 11.9% at $100-499, 6.2% at $500-999, 4.0% at $1,000-1999, 3.8% at $2,000-4999, and 4.2% above $5,000. Conclusion, the cost of an accommodation should not be an employment determining factor.

The types of personal computer adaptive devices, programs and aids for a person with disabilities is practically unlimited. For a person who may be blind or low vision, for a person in a wheelchair or orthopedically impaired, for a person who is deaf or speech impaired, or for a person with a learning disability, there are solutions to aid their education, personal living, or employment opportunities. The IBM National Support Center for Persons with Disabilities has compiled a disability resources file that contains over 600 products, over 500 vendors and over 700 disability agencies and groups. These disability resource reports are available via a toll free number, (800) 426-2133, from 8:15 am to 5:15 p.m. EST, Monday thru Friday of each week. Since its formation in December of 1985, the Center has responded to over 15,000 inquiries from all over the world.

In addition to the disability response line, the National Support Center conducts disability briefings in Atlanta for employers, educators, rehabilitation professionals, government officials, and others who have an interest in persons with disabilities. In 1987, the Center began a series of Executive Awareness Programs to take these briefings outside Atlanta, working through the local IBM branch offices to raise the level of awareness of the same groups of people.

The challenge today is to raise the level of awareness of employers, educators, job placement counselors, government officials, rehabilitation professionals and the general public as to the capabilities of persons with disabilities. The technology is available to allow them to obtain quality education, have a normal life style with self-respect and to have equal opportunity to competitive employment positions.

Awareness alone, however, is not enough. We need to begin opening doors...doors to quality training...doors to availability to required technology...doors to competitive job opportunities without discrimination...doors to that dream that we all have: The door to independence, self-respect, meaningful employment opportunities and the ability to use our God-given talents to be self-sustaining in our every day life.

We are on the threshold but we need your help in conquering the inequities that exist for this part of our population in this wonderful country of ours today. Can I count on you???

* * * * * * * * * *

MULTI-PURPOSE JOB INFORMATION SYSTEM : DESIGN AND APPLICATIONS

Robert G. Pajer,
Chief, Validation and Analysis Staff
U.S. Drug Enforcement Administration

Abstract

The workshop explored how to design a job information system, provided a demonstration of its capabilities and considered relevant applications of an automated job information system to workshop participants.

The Drug Enforcement Administration Job Information System (DEAJIS), a human resource management data base system, was initially described and then discussed as a model to maintain the job relatedness of personnel management functions such as employee training, career development and performance appraisal. The latter was the focus as we examine the utility of a fully automated, operational, behaviorally-based performance appraisal

program with on-line data entry, monitoring (editing), analysis and report capabilities to areas of personnel management.

Participants shared experiences associated with the development and implementation of job information systems and considered how an automated system such as the described model can be used to meet particular personnel management needs.

Design Overview

The Drug Enforcement Administration (DEA) has recently completed a comprehensive job analysis of the Special Agent Criminal Investigation occupational series. The results of the job analysis are used to validate aspects of DEA personnel management and to establish ongoing support for employee development, performance appraisal and position management. DEA has established an automated, on-line job information system (DEAJIS), a mainframe data base system to maximize the benefits derived form its multi-purpose job analysis. DEAJIS presently supports four major objectives: document the job analysis and data collection records, entry and analysis of performance appraisals, inquiry against job information and linkages to other DEA personnel/human resource management system. DEAJIS has the following on-line capabilities:

* Store and support the periodic updating of outputs of the Special Agent job analysis.
* Enable users to compose, edit and compare job titles.
* Identify qualified employees for internal recruitment.
* Provide records of the job analysis data collection process.
* Provide reports needed to support personnel management decision making.
* Support entry and analysis of performance appraisals, test development, training needs identification and career development planning.

These functions are accessed through a user friendly, menu-based system.

DEAJIS is organized into two major subsystems. One subsystem supports research and development of improved personnel and the other supports personnel operation.

The research and development subsystem provides support to personnel operations. Three functions associated with the research and development subsystem are:

1. Selection Validation - this functions assists in the validation of selection procedures by supporting the statistical analyses of quantitative data associated with the job information.

2. Query - the Query function allows for exploration of DEAJIS information is in a very efficient manner. The user may specify what categories of job information are to be explored and the system identifies the relevant linkages (e.g., what KSA's are associated with a particular category of work).

3. Systems Development - this component represents the intent of the system to support research into new applications. The subsystem is being designed to facilitate the addition of new functions as they are needed.

The purpose of the personnel operations subsystem is to provide automated support for the day-to-day implementation of the job analysis and other personnel functions. Specific functions include updating of job information files with newly validated job information, the development of and maintenance of job titles, the identification of candidates for stages of career advancement, the preparation of performance appraisal plans and entry, monitoring (editing) and analysis of appraisal ratings, the assessment of training needs and the preparation of crediting plans.

System Interfaces

DEAJIS incorporates the following features to enhance its utility:

* Several locations for outputting DEAJIS products are under menu control such as the remote terminal and the laser printer.

* A Help function has been designed to allow users to locate any job information by entering a key word or phrase.

* Production reports have been designed to allow users to locate any job information by entering a key word or phrase.

* The structure of the DEAJIS menus and the language used to identify its functions reflect

the way personnel management operations are actually organized in the Agency.

This workshop was developed by Gary L. Musicante, Senior Psychologist, U.S. Drug Enforcement Administration, Washington, D.C.

* * * * * * * * * *

TEST SECURITY VS. APPLICANT RIGHTS

George Rost, Assistant Chief Examiner
City of Los Angeles

A. What are the concerns of the applicant
   - -Appropriate selection devices
   - -Correctness of material used
   - -Equal treatment for all

   What are the concerns of Personnel
   - -same as above plus
   - -security of materials

B. City of Los Angeles - eras of change
   - -The 1930's - corruption and reform
   - -1939-1972 - open process
   - -1972-1976 - conflicts
   - -1976-Present - changes in protests and reviews

C. Changes to Applicants' review rights
   - -CSC concern about validation
   - -validation vs. review
   - -proposal
   - -union reaction
   - -final action
     - expert review include union nominee
     - candidate can protest test administration and job relatedness
   - -reaction and acceptance

D. Changes to Applicants' Protest Procedure
   - -CSC concern about time delays and frivolous protest (written and interview)
   - -solution for written test protest - due process
     - support for proposed change
     - reviewed by staff and subject matter experts
     - mutually agreed on expert panel
     - GM accepts recommendation - final decision
   - -solution for interview protests - timing and definitions

correct time to protest - 48 hours
support for protest
defining what can be considered
Feedback

E. Does it work
Review
-we get better review than before
-similar number of changes made than under old protest system
Protests
-much more orderly and much simpler
-saves time

Applicant Acceptance - very good

* * * * * * * * * *

THE VEIL OF SECRECY

Amy Eagan and Thomas Davis
Columbus, Ohio Civil Service Commission

OLD METHOD

Entry level:

Inspection Period: Ten calendar days immediately following written notification of final grade and position on the list.

Examinees may inspect their answer sheets for possible grading errors by comparing them with a keyed answer sheet provided by the Commission.

No examinee may see the test materials after an examination.

Promotional Level:

Test Site: Candidates are permitted to see the correct answer key and a test booklet immediately following the exam to appeal specific multiple-choice items.

Appeal Period: Five calendar days from the test date. Candidates are not permitted to see the test booklet during this time.

Any item can be appealed at this time. Each appeal will be investigated and a decision will be rendered by the Executive Secretary within 30 days.

Inspection Period: Same as for entry level.

Work Sample: Candidates are given the total number of points possible for each problem and the total number of points they received for each problem.

Candidates may not inspect their test booklets or answer keys at the test site or during the appeal and inspection periods.
Oral Boards: Candidates may see their score broken down into two areas: Content & Style.

Candidates may also listen to the audio tape of their interview and/or watch the tutorial tape for Phase IV that was shown before the examination. The tape gave an example of both a good and a bad oral presentation.

NEW METHOD: (proposed)

Entry Level: Same as old method

Promotional Level:

Test Site: Candidates will be permitted to see the answer key and their test booklet immediately following the exam. The candidate's answer sheet will be collected prior to the release of the correct answer key. Candidates, however, will have been instructed at the test site that they are permitted to write and mark their answers in their test booklets.

Subsequent Appeals: Three Civil Service Commission work days following the examination in which the candidates may see the "correct" answer key and an unmarked test booklet.

Appealable Items: Multiple-Choice test items can only be appealed for the following reasons:

1. No correct alternative
2. Multiple correct alternatives
3. Incorrectly keyed alternatives
4. Keyed alternative conflicts with one or more knowledge source

Ambiguous appeals will be dismissed.

Inspection Period: same as old method.

Work Sample: Candidates may see the answer key and the test questions at the test site in order to formulate appeals.

Appeal Period: Three Civil Service Commission work days following the examination.

Appealable Grounds:

1. Examinee's response is correct and is not listed in the keyed response set.
2. The keyed response set is not correct or conflicts with established policies and/or procedures.

Inspection Period: Ten calendar days immediately following the exam. Candidates may see their response sheet and their score sheets.

Oral Boards: (We have not yet decided what actions or changes we will take in this area.)

A SURVEY OF APPEALS PROCEDURES AT SELECTED U.S. CITIES

CITIES CONTACTED FOR RESEARCH

Rank (by pop)	City	Population	Police Officers
5	Philadelphia, PA	1,646,713	6,868
9	Phoenix, AZ	853,266	1,725
12	Baltimore, MD	763,570	2,976
18	Milwaukee, WI	620,811	1,978
19	Jacksonville, FL	577,971	963
21	Columbus, OH	566,114	1,224
22	New Orleans, LA	559,101	1,305
23	Cleveland, OH	546,543	1,701
24	Denver, CO	504,588	1,310
25	Seattle, WA	488,474	1,063
33	Pittsburgh, PA	402,583	1,128
38	Cincinnati, OH	370,481	875

DATA FROM THE 1988 INFORMATION PLEASE ALMANAC

Please note that other cities, besides those similar to Columbus, were also contacted for additional information.

Several factors were not taken into consideration in the selection of the above cities to use for this research:

- -growth since 1980 (data from 1980 census)
- -relative crime rates
- -make-up of the population (wealth, race, etc.)
- -legal differences (definitions)
- -geographical (land) size of the city
- -size/influence of the local fire or police unions
- -accuracy of data source
- -several other related factors.

This study considered the tests for the Police and Fire Departments. Only the number of Police Officers was readily available and used for comparison. Fire Department figures could not be obtained.

SUMMARY OF RESULTS

1.	3 out of 11	have separate people or departments for Police and Fire testing.
2.	4 out of 8	use consultants to do job analyses.
3.	9 out of 11	have 100% multiple choice entry level tests.
4.	2 out of 11	have 100% multiple choice promotional tests.
5.	4 out of 11	have work sample/essay portion of tests.
6.	8 out of 11	have oral exam/interview portion of tests.
7.	7 out of 11	use different types of promotional tests for different ranks within the same department.
8.	4 out of 11	allow appeals for entry level tests.
9.	1 hour to 30 days	is the range of time allowed for appeals.
10.	2 out of 10	have specific preset grounds for submitting appeals.
11.	2 out of 10	use consultants exclusively to write exams.
12.	6 out of 10	use consultants to write exams.
13.	3 our of 9	use behavioral anchored rating scales.
14.	3 out of 6	who do not use behavioral anchored rating scales, use scales with general definitions such as "excellent" or "acceptable".
15.	1 out of 3	use consultants to develop the behavioral anchors for the rating scale
16.	2 out of 8	have people other than Police and Fire Department officials on the oral board.
17.	5 out of 8	have oral board members exclusively from other jurisdictions.
18.	2 to 6	is the range of average number of oral board members (when specified).
19.	4 out of 8	always have exactly 3 oral board members.
20.	4 out of 8	give specific scales to be rated to the candidate prior to the oral exam.

21.	30 mins. to 45 min.	is the range of average time for oral exams (when specified)
22.	0 out of 11	have on-site grading.
23.	5 out of 11	allow the candidate to use the answer sheet to formulate appeals.
24.	8 out of 11	allow the candidate to use the key and a test booklet to formulate appeals.
25.	6 out of 8	allow the candidate to use a keyed test booklet to formulate appeals.
26.	5 out of 7	use both double keying and elimination to correct valid appeals.
27.	4 out of 7	give the candidate the dimensions on which the candidate was rated, and the ratings, after the oral exam.
28.	3 out of 8	allow candidates to see rater's comments after the oral exam.
29.	3 out of 8	use audio tape to record oral exams
30.	4 out of 8	use video tape to record oral exams.
31.	1 out of 8	use both audio and video tape to record oral exams.
32.	2 out of 8	do not use oral exams at all.
33.	1 out of 8	allow candidates to review audio tape to formulate appeals.
34.	2 out of 4	allow candidates to review video tape to formulate appeals.
35.	4 out of 8	allow appeals on non-uniformed exams.
36.	6 out of 8	have different appeal procedures for uniformed and non-uniformed exams.
37.	2 out of 7	use z-scores to convert scores (when specified).
38.	2 out of 11	have one person that rules on the appeals.
39.	2 out of 7	allow appeals of rulings on original appeals.

The first number is the number of cities that meet the condition. The second number is the total number of cities that responded to that particular question, or the total number of cities to which the condition applied.

* * * * * * * * * * *

TEST SECURITY, APPLICANT RIGHTS AND THE CANDIDATE REVIEW PROCESS

Paul D. Kaiser, Principal Examiner
N.Y.S. Dept. of Civil Service

REVIEW PROCEDURE DESCRIPTIONS:
PRE-RATING REVIEW
PRIOR APPROVAL REVIEW
POST-RATING REVIEW
COMPUTATIONAL REVIEW

LEGAL BASIS: THE NEW YORK STATE CIVIL SERVICE COMMISSION RULES STATE THAT THE INTENT OF THE REVIEW PROCESS IS TO CONSIDER CANDIDATE OBJECTIONS THAT, "CLEARLY DEMONSTRATE A MANIFEST MATERIAL ERROR OR MISTAKE APPEARING IN A RATING KEY OR SCALE OR IN THE APPLICATION OF SUCH KEY OR SCALE TO CANDIDATE TEST PAPERS OR OTHER RECORDS OF EXAMINATION PERFORMANCE OR ELIGIBILITY FOR APPOINTMENT AND ONLY IF SUCH ERROR MISTAKE AFFECTS THE LEGALITY OR RELATIVE STANDING OF CANDIDATES."

POLICY CONSIDERATION: THE DEPARTMENT OF CIVIL SERVICE POLICY MANUAL STATES THAT, "THE PURPOSE OF OPENING TEST MATERIAL TO CANDIDATE REVIEW IS TO DEMONSTRATE, AFFIRM AND SUPPORT THE PRINCIPLE OF FAIR AND OPEN COMPETITION FOR CIVIL SERVICE EMPLOYMENT. THE DEPARTMENT PERMITS CANDIDATE REVIEW TO THE EXTENT THAT SUCH REVIEW DOES NOT CONFLICT WITH THE REASONABLE REQUIREMENTS OF TEST SECURITY."

PRE-RATING REVIEW - Under this procedure candidates are allowed to inspect the test questions and the Department's tentative answer keys and to submit objections to the proposed key. Candidates do not see which choices they selected. However, candidates are allowed to bring books and other reference materials with them to the review center. Candidates may not bring a consultant or send a representative in their place. Pre-rating reviews are usually conducted the Saturday following the announced examination date. This review procedure is used most often and thus generates the greatest number of candidate objections. Any changes in the answer key resulting from this review affects all candidates.

This procedure is used only for multiple choice or short answer written tests.

PRIOR APPROVAL - This is a no review or appeal process where keys are confirmed by the Civil Service Commission prior to the administration of the examination. This procedure is used when test questions have been tried and proven, that is, have been

though the review process several times and the Commission has repeatedly confirmed the answer keys. Prior approved status may also be granted when the items set problems which contain, within themselves, sufficient information to completely determine the best response, e.g., arithmetic, spelling, etc.

This procedure is used only for multiple choice or short answer written tests.

POST-RATING REVIEW - This procedure is in effect for all examinations not covered by prior approval or pre-rating review conditions. Candidates are permitted to appeal after the eligible list has been established (hence, post-rating). Under this procedure, candidates meeting certain conditions are permitted to inspect the test questions, the "final" rating key and their own test papers and submit objections against the rating key. It also includes a computational review. (see below).

This procedure is used primarily when candidates are rated against a scale (orals, T&E's, essays, etc.) and is infrequently used for multiple choice or short answer tests.

COMPUTATIONAL REVIEW - Under this procedure, candidates may inspect their answer paper, the final rating key, and any scoring table or scoring formulas used in converting or transforming their scores and may submit objections against the application of the rating key to their paper. In essence, the computational review is a check by the candidates to see that his or her examination paper was scored correctly.

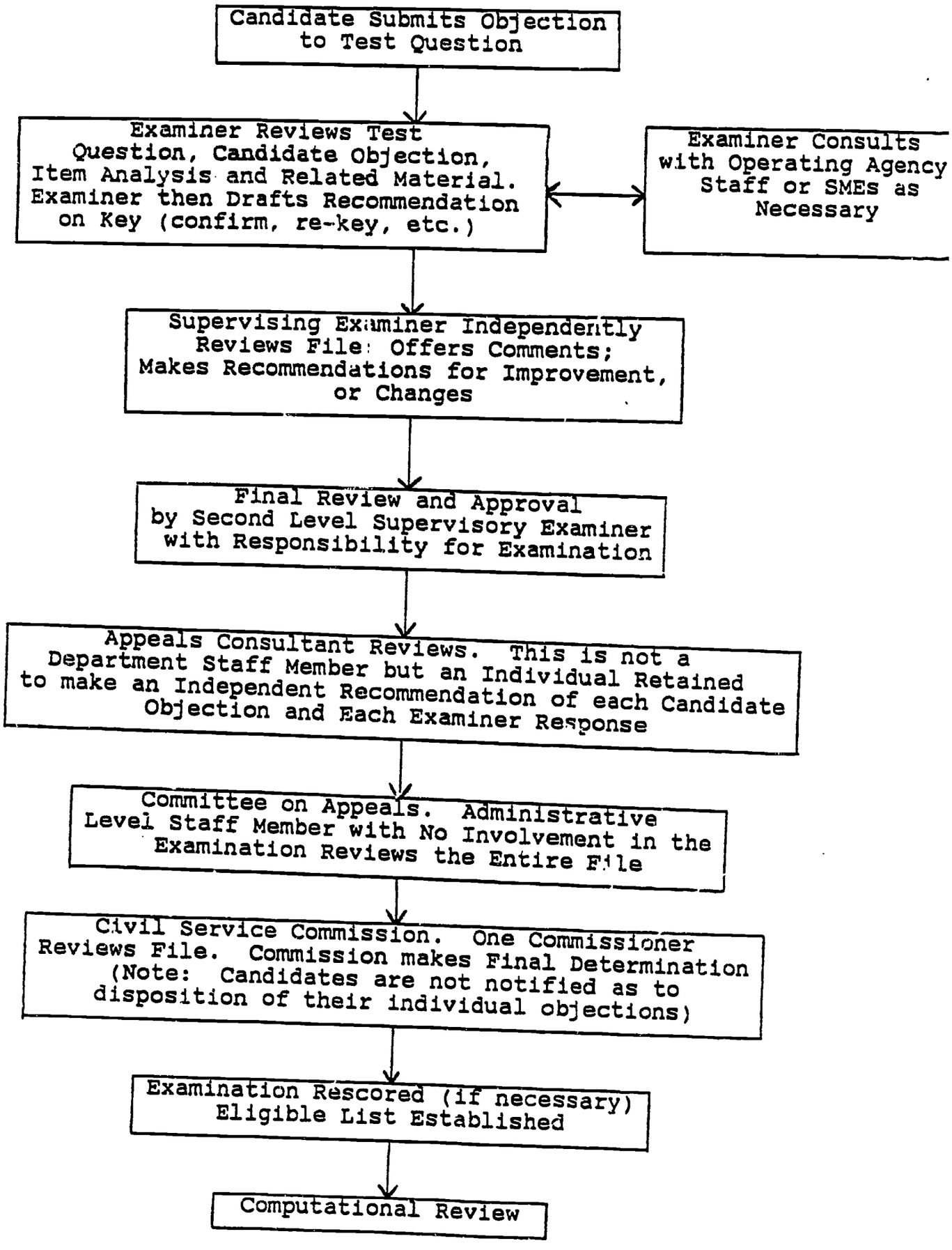
This procedure is allowed for all tests.

Types of Review

Types of Tests	Prior Approval	Pre-Rating Review	Post-Rating Review	Computational Review
Oral test	No	No	Yes	Yes
T & E	No	No	Yes	Yes
Written	Yes	Yes	Yes	Yes
Cont. Rec.	Yes	No	No	Yes

Written Examination Review Process

Flow Chart



EFFECTS OF THE CANDIDATE REVIEW PROCESS
SURVEY OF TESTING DIVISION STAFF

1. Which statement best describes your opinion concerning the "fairness" of this Department's examination review process with respect to candidates' rights and interests?

#	
8	A. Our review process is not as "fair" as it might be.
26	B. Our review process is as "fair" as it can be and generally should remain as is.
18	C. Our review process is more than "fair" to the candidates. We should take measures to limit the process.

2. Which statement best describes your opinion of the benefits to the candidates of the examination review process?

#	
6	A. The candidate generally do not benefit from the review process.
21	B. The candidates slightly benefit as a result of the review process.
21	C. The candidates moderately benefit as a result of the review process.
6	D. The candidates greatly benefit as a result of the review process.

3. Which statement best describes your opinion concerning the effects of the appeals process upon test security considerations?

#	
27	A. The appeals process does not compromise the security of our exam materials in any serious way.
21	B. The appeals process does have a slight compromising effect on the security of the exam materials.
6	C. The appeals process greatly compromises the security of our exam materials.

4. How many key/score or qualifications changes would you say occur in your examinations as a result of the candidate review process?

#	
48	A. Less than 3 changes per exam series
6	B. Between 3 and 5 changes per exam series
0	C. More than 5 changes per exam series (please specify) _____

5. Thinking back over the answer or score changes that you might have made as a result of the appeals process, what percent of these changes would you have likely made on your own (e.g., through review of statistical results or post test meetings with SME's, etc), without the candidates bringing the issue(s) to your specific attention?

#		
9	A.	None
1	B.	Between 1 - 5%
3	C.	Between 5 - 10%
3	D.	Between 10 - 25%
9	E.	Between 25 - 50%
15	F.	Between 50 - 75%
12	G.	Between 75 - 100%

6. Which statement best describes your opinion of the value of the appeals process with response to improving the QUALITY of our examinations?

#		
6	A.	No value
18	B.	Some, but little value
30	C.	Of some moderate value
6	D.	Of great value

7. How much time do you spend dealing with candidate objections submitted through the review process appeals during any given year? (this includes not only responding to appeals but also time spent in administering the process; e.g., "paperwork")

#		
21	A.	Less than 5% of the unit's time
24	B.	Between 5% - 10% of the unit's time
6	C.	Between 10% - 20% of the unit's time
6	D.	Between 20% - 30% of the unit's time
0	E.	Other (please indicate percentage) _____

8. In your estimation, what is the effect of the appeals process on the timing of the establishment of eligible lists?

#		
18	A.	No effect
9	B.	Slows down establishment by 1 - 2 weeks
6	C.	Slows down establishment by 2 - 4 weeks
6	D.	Slows down establishment by 4 - 6 weeks
12	E.	Slows down establishment by 6 - 8 weeks
0	F.	Slows down establishment by more than 8 weeks (please specify) _____

9. Which statement best describes your opinion of the benefits to our Department of the examination review process?

#	
0	A. Our Department generally does not benefit from the review process.
30	B. Our Department slightly benefits as a result from the review process.
21	C. Our Department moderately benefits as a result of the review process.
6	D. Our Department greatly benefits as a result of the review process.

PROCEDURE FOR EVALUATION OF TRAINING AND EXPERIENCE APPEALS

I. T & E Appeals

A. Applicants commonly protest that:

1. The rating scale is in error
   - the wrong training/experience factors were considered
   - too few training/experience factors were considered
   - scale was improperly developed
   - weighting of training/experience is wrong

2. The application of the rating scale is in error
   - subject matter experts and/or raters were not properly briefed or qualified
   - insufficient credit give to certain kind(s) of experience
   - level/scope/relevance of candidate's experience misinterpreted by rater
   - applicant knows someone with the "same" experience who received a better score

3. The rating of training and experience was an inappropriate examination
   - should have been a written/oral test

4. The weighting of the T & E portion of the examination was inappropriate
   - should/should not have been weighted
   - should/should not have been qualifying

B. The appeal process[1]

1. Any one or more of the above factors may constitute grounds for an appeal. Although the basis for sustaining an appeal should, ordinarily, be limited to a demonstration of manifest error, standards for developing training and experience evaluations found in the Department's T & E Manual will provide further guidance to staff on what can and should be defended.

2. The assumptions which underlie the appeal process are as follows:

   a. The information available in the examination folder for review includes the job analysis information, the rating scale (including documentation of its development and justification), the scoring procedure, and subject matter expert documentation.

   b. Every reasonable attempt has been made by the responsible Staffing or Testing Representatives to avoid an appeal. This would include negotiation with agencies on minimum qualifications, test plan, and test format; explanation of rationale behind minimum qualifications, test plan, crediting plan, rating scale (as appropriate) to candidates or agencies; re-review of application or supplemental forms to assure that a correct and reasonable determination has been made; and an explanation of that determination.

   c. The training and experience examination was developed, insofar as possible, in accordance with the Uniform Guidelines for Employee Selection Procedures.

C. The T & E Appeals Procedure is as follows:

1. As stated on the XD-230 or XD-230.1 Notification of Examination Results form, the candidate is allowed ten business days after the postmark date of the notice of results to request review of the marking of his/her papers. This request may be in the form of a letter or a telephone call. Telephone inquiries should be handled by the responsible Staffing Representative. Every reasonable attempt should be made to answer the candidate's questions about the examination. If the candidate is not satisfied or a complete explanation is not possible over the telephone, the candidate should

be instructed to send in his/her questions or objections in writing.

a. The candidate may request information concerning his/her score. This might include a request for reevaluation or for an explanation of the rating procedure.

b. The candidate might also introduce additional (not previously submitted) information for evaluation. Any additional information must be disregarded since it would be received after the eligible list has been established or--in the case of multi-part examinations--after the applications have been evaluated. To accept it would require a reevaluation of the candidate and could potentially give him/her an unfair advantage over the other candidates.

c. The candidate might simply request an opportunity to review the marking of his/her papers. In this case, the candidate should be informed of the appeal procedure as in C.2.c., and the procedure outlined in C.3. should be followed.

2. The Staffing Services Representative responsible for the examination in question responds to the candidate's inquiry.

a. If the inquiry is general and the Staffing Services Representative feels it can be handled by a narrative-type explanation, he/she should send the candidate a standardized or an individualized letter containing, as appropriate, an explanation of the rating scale and the crediting system.

b. If the candidate's question is on how his/her score was arrived at, an explanation of the credit given or not given for the candidate's experience should be included.

c. In addition to the explanations given in the letter, the appeal procedure should be outlined as follows: The candidate should be informed of his/her right to appeal and that the only basis for sustaining an appeal is the proof of the occurrence of manifest error. Manifest error is defined as an actual error or mistake in any aspect of the examination process. The burden of proving manifest error rests with the appellant. The candidate should be given

ten business days from the postmark date of the explanatory letter to request a review of the marking of his/her papers. The candidate should be informed that requesting a review constitutes the first step of an appeal and will be treated as such.

3. The point at which a request to review the marking of the candidate's papers is received will be considered the first step in the formal appeal process.

   a. A standardized letter should be written to be sent to all candidates who request a review of the marking of their papers. The candidate should be informed that his/her request is being considered as an appeal and that he/she has 14 business days from the postmark on the envelope to send in complete objections in support of an appeal. The letter should state that additional information concerning training and/or experience not previously submitted in the original application and/or supplemental application will <u>not</u> be considered.

   b. When a request for review is received, the following examination materials should be copied and sent to the candidate (along with the letter described in 3.2.):
      - the rating scale used for the examination
      - definitions of terms used (if necessary)
      - a photocopy of the candidate's rating sheet(s)
      - a photocopy of the candidate's supplemental application form (if used)
      - an explanation of how credit was applied
      - a photocopy of the candidate's original application would not routinely be included, but would be made available is requested

   c. When an appeal has been made, the entire administration of the examination is open to review.

4. When objections are received, the responsible Staffing Representative responds by writing a memorandum to the Commission, addressing the points of appeal. The memorandum is then forwarded through the Division Director's Office to a Consultant on Appeals. Included with the memorandum is an updated "final" letter to the candidate for the President's signature, based on the

Staffing Representative's recommendation to sustain or dismiss the appeal.

a. The material submitted to the Consultant in response to the appeal should include at least:
   -the memorandum described above
   -the original application and supplemental application (if used) submitted by the applicant
   -a copy of the rating forms and scales
   -an explanation of how the forms and scales were applied--generally, and in this case
   -copies of relevant correspondence with the candidate
   -in some cases, it may be useful to provide applications and rating sheets of other candidates to provide the reviewer with examples of higher and lower quality experience
   -staffs recommendation to sustain or dismiss the appeal

   Staff should be aware that certain kinds of actions in response to an individual's T & E appeals will have an effect on the entire examination. Sustaining such an appeal may even result in the invalidation of the procedure used to evaluate remedies for their potential to disturb the entire examination process.

   b. The Consultant on Appeals will review all material, make a recommendation concerning the disposition of the case, and forward all of the material to the Commission's Committee on Appeals.

   c. The Commission's Committee on Appeals reviews the material, makes a determination, and the item is formally considered at the next regular Commission meeting as an examination appeal. As with all other examination appeals, that candidate normally will to be allowed to argue his/her case or present additional information before the Commission, since the primary review of the record is made by the Consultant and the Committee based on the full written record.

   d. Upon conclusion of its review, the Commission either forwards the letter provided by

the Staffing Representative or requests a revision. The entire file is returned to the Staffing Representative through the Division Director's Office. The Staffing Representative notes in the examination folder that the appeal is completed and forwards the appeal materials to Central Files.

* * * * * * * * * * *

SELECTION OF POLICE MANAGERS IN AN ENVIRONMENT HOSTILE TO THE ASSESSMENT CENTER

Patrick T. Maher, Principal Associate
Personnel & Organization Development Consultants, Inc.
La Palma, California

The examination took place in a municipal police department with a department staff of 150 sworn and non-sworn. The department had three assistant chiefs of police, who reported to the chief.

During the job analysis, hostility to the assessment center was noted. While the chief and the police commission as a whole, were open to an assessment center, an assistant city manager, some candidates, at least one assistant chief, and some individual police commissioners were opposed to or leery of it.

Several months prior to the examination, several lieutenants, including some taking the examination, had conducted a research project that concluded that the assessment center was "not producing the desired results."

The department had used the assessment center for examinations for lieutenant and sergeant, and for career development process. Each assessment center was conducted differently and these experiences had created some dissatisfaction with and concern about assessment centers. Some specific concerns included: The validity of the assessment center as the sole criterion for ranking or selection; inconsistent ratings of candidates among assessment centers; lack of or inadequate departmental input into the promotional process.

It is clear, however, that these problems related to the assessment procedures rather than the assessment center method itself. For example, lack of departmental input was inappropriate. Thus,

the assessment center method was improperly blamed for defects in total examination design. This important point should be considered whenever analyzing dissatisfaction with the assessment center method.

It must be remembered that no examination device is perfect, and that criticisms leveled at the assessment center method have also been leveled at other assessment procedures. Thus, the abandonment of the proven assessment procedure without careful consideration of the facts will only result in the adoption of other methods that will also eventually produce dissatisfaction. Indeed, at one time many agencies unrealistically adopted the assessment center as a panacea for problems found in other assessment procedures. As with all such expectations, the cure became the curse.

It is important to note that the assessment center method has been proven to be psychometrically sound and a number of courts have recommended the assessment center as an alternative to procedures being challenged in Title VII cases.

Another important consideration is that the assessment center method is frequently misused in the public sector. Many procedures identified as assessment centers do not comply with the Standards and Ethical Considerations for the Assessment Center Method (Standards). Therefore, the process being identified as an assessment center must be carefully analyzed to see if it actually is one before experience with the procedure should be the basis for rejecting the assessment center method. In the department we have been discussing, some of the "assessment centers" did not conform even superficially with the Standards.

Because there was such reservation or dissatisfaction with the assessment center method in this police department, it was recommended that it not be used in this examination. Instead, it was decided that the small number of candidates, all of whom were internal to the department, made other assessment procedures viable.

As a part of the final decision on examination design, project staff met with all candidates and discussed their concerns about the various proposals and issues. To the greatest extent possible, their doubts were addressed and resolved. It was this consultative process, more than anything else, that probably accounted for the general candidate acceptance.

To evaluate the candidates, a rating panel consisting of a chief of police from outside the county, a police commissioner, and a citizen from the community was used.

As a direct result of the meeting with the candidates, the chief of police served as an ex officio member of the rating panel. In

this role, he provided additional perspectives of each candidate's actual on-the-job performance, and was a resource to determine if job experiences claimed by the candidates were accurate. His presence provided departmental input, although he did not rate candidates.

An oral presentation was used to measure performance under simulated job conditions. In this exercise, candidates were given background information on an officer-involved shooting scenario. After preparation, candidates gave an uninterrupted 5-minute oral presentation to the rating panel, which served as the city council. During this presentation, the candidate had to summarize the incident and indicate the department's position on the shooting (i.e., justified or not justified).

The panel then asked questions to determine how well the participants would respond. Suggested questions, prepared ahead of time, were designed so that no matter what position a candidate took, the panel could ask questions hostile to the candidate's position.

The assessment center method has long recognized the background interview as a viable means of integrating information from outside the assessment center into the judgement of critical skills. Often, behavioral-based or situational interviews that have recently come into use are really nothing more than an adoption or adaption of the assessment center's background interview.

Prior to the interview, each candidate completed an extensive questionnaire that covered not only job experience, but other experiences that might reveal relevant behaviors in the dimensions being assessed (e.g., community services or activities, military service, specialized training, etc.).

This questionnaire was then reviewed and specific questions in each dimension for each candidate prepared by the consulting staff.

The rating panel asked these prepared questions as "primary questions" and then asked any follow-up questions it deemed necessary to obtain relevant behavior.

Initially, the rating panel only obtained and documented behaviors (responses). Once all of the candidates had been interviewed, the rating panel reviewed the recorded responses and independently rated candidates in each dimension.

Once independent ratings were assigned, the panel met for an integration discussion, as is typical to a proper assessment center. If the scores for all three raters were identical, no discussion was conducted unless one of the raters felt that

something needed consideration. If even one rater had only one score difference, discussion was conducted to determine why there was a disagreement. While unanimity was sought, it was not mandated. The discussion's purpose was to determine if there was a reason for the difference.

The chief was present during this process to again provide additional perspectives about on-the-job performance for the rating panel. The panel had the option of changing the scores based on the chief's input or keeping its scores the same. Thus, while the chief was available for providing information, he did not have any special power or authority to change scores.

Although this procedure used the psychometric aspect of the assessment center method as well as an assessment center simulation exercise, it did not constitute the full assessment center method. In addition, it provided departmental input by having the chief serve as an additional information resource and by making on-the-job performance information available through personnel files.

While it is always difficult to assess test satisfaction, several factors would indicate that there was general satisfaction with the test.

First, the chief appointed the first-ranked candidate from the list. Then, several months later, a newly-appointed chief promoted two candidates to assistant chief in rank order.

After completion of the examination, candidates were asked to evaluate the process. They rated the extent to which they felt that the simulation exercise and the background interview were job related. On a 5-point scale, both received a mean rating of 4.29. In addition, they were also asked to indicate the extent to which they felt that this test was better than or worse than an assessment center. A "5" meant that the process was better than an assessment center. The mean rating for this scale was 4.43, with five of the candidates giving a "5" rating. Based on these ratings, we concluded that the candidates were satisfied with this testing process.

Based on these and other facts, we concluded that the testing process for assistant chief enjoyed broad departmental support and acceptance across all levels.

In addition to the queries about job relatedness and comparison with assessment centers, candidates were asked other questions about the process. They were asked to rate each candidate as to how well they thought he would perform if promoted to assistant chief by ranking the best performer first, the second best performer second, and so on. They were then asked to rank-order the candidates as to how they thought they would score on the

test, regardless of how qualified they might be on the job. Interestingly, while the candidates rated the test components as being "job related" and better than an assessment center, they felt that test performance would be different from job performance.

While we were unable to determine why this dichotomy existed, we concluded that the candidates did not trust the test to accurately measure relative ability, even though they felt that it was job related. Thus, their complaints about the assessment center results being different than their perception of true performance are not limited to the assessment center.

The two incumbent assistant chiefs were also asked to rank-order the candidates as to how they thought they would perform on the test. There was some variance between the two that showed that they did not agree on who would be the best test performer. Therefore, we concluded that any disenchantment with the results would apply to any assessment procedure.

There seems to exist a concern, bordering in some cases on paranoia, about the use of the assessment center. Yet, the assessment center was itself first viewed as an alternative to other selection procedures.

As this examination shows, an assessment center does not have to be used if it is not amenable to a given testing situation. Furthermore, using different assessment procedures merely because of pronouncements of dissatisfaction from candidates may not result in greater acceptance. This assessment procedure was accepted because we listened to the specific concerns of all and made a conscientious effort to address as many as possible, under myriad constraints. Had there been time, we believe we could have rehabilitated the assessment center process.

While the assessment center remains a viable assessment procedure, there need be no concern about using or finding alternatives to assessment center. It is up to the psychometrician to decide when and how it is best used in a given testing situation. Panaceas do not exist.

Selected References

Jaffe, C.L., & Frank, F.D., Interviews conducted at assessment centers. A guide for training managers. Bubuque, WI: Kendall/Hunt Publishing Company, 1976.

Schmidt, F.L., Caplan, J.R., Bemis, S.E., Decuir, D., Dunn, L., & Antone, L. The behavioral consistency method of unassembled examining. (TM-79-21) Washington, D.C., U.S. Office of Personnel Management, Personnel Research and Development Center, November 1979.

Firefighters Institute v City of St. Louis, (1977) 14 FEP Cases 1486

Harless v Duck (1980) 22 FEP Cases 1073

Williams v City & County of San Francisco (197() 22 FEP Cases 1241

* * * * * * * * * *

EMPLOYEE OPINIONS OF FOUR PROMOTIONAL EXAMINATION MODES

Joel P. Wiesen, Director
Applied Personnel Research
Newton, Massachusetts

Summary

The state of Connecticut commissioned a program evaluation of its new (seven year old) merit board system of civil service promotional examination which is known as "MPS". MPS is basically a committee-based unassembled examination system. It is one of four modes of promotional civil service examination in Connecticut. The evaluation was undertaken against a backdrop of some amount of negative opinion about MPS, and growing pressure on the legislature to make promotional examinations subject to collective bargaining. The program evaluation included a survey of opinions and attitudes of Connecticut civil service and exempt employees and managers toward all four examination modes and toward promotional examinations in general. The program evaluation was the basis for formulating recommendations for improving the state's merit board promotion system.

The program evaluation began with the original goals for MPS, for example: timeliness, reducing provisional appointments, performance, allowing agencies a more substantive role, giving credit for job performance, and reducing the examination workload. Accomplishments in each of these areas were summarized.

Since attitudes and perceptions were a major issue, a survey was undertaken. The survey replicated and expanded on one conducted about 5 years ago. Opinions were probed in areas such as practicality, and adherence of each examination mode to the merit system principles. The four specific modes of promotional examination considered are: written, T&E, oral and MPS. The responses were considered in light of self-identification (bio-

data); for example, the responses of managers, non-managers, union members and non-union members were compared.

The attributes rated as most important for a civil service examination were: fairness to all applicants, and selecting the best qualified applicants. Safeguards to abuse was rated third (substantially higher than adequacy of appeal procedures).

MPS was the most fully accepted of the four examination modes. Overall satisfaction with MPS was high; 53% of employees who applied for but were not appointed reported being satisfied with MPS.

The survey also attempted to measure knowledge about this relatively new examination mode by use of a true-false test. There were some crucial gaps in knowledge about MPS among each group of employees (e.g., managers, supervisors), but particularly among non-supervisory employees.

As a result of the survey and the larger program evaluation a number of program changes were recommended in several areas, including:

- o publicity and training
- o simplification
- o announcing examination areas (KSAPs)
- o reliability (especially across merit boards)
- o fairness
- o feedback to applicants on ratings
- o appeals of MPS ratings
- o degree of position specificity of examinations
- o additional research needs
- o live audits (in addition to post audits)
- o staffing level guidelines for MPS functions
- o need for a formal, written validation report

Beyond these areas which are specific to MPS, several changes were recommended which relate to the overall merit system, such as:

- o reevaluate and clarify the State's policy on promotion to filled positions
- o address special needs arising from a lenient certification law

The State of Connecticut is now in the process of implementing many of these recommendations.

Note: A limited number of copies of the full report are available from the author.

* * * * * * * * * * *

JOB SATISFACTION IN THE FEDERAL WORK FORCE

Paul van Rijn
U.S. Merit Systems Protection Board

This paper describes the results of a survey of job satisfaction among Federal employees that was conducted by the U.S. Merit Systems Protection Board during 1986. A disproportionately stratified random sample of 21,620 employees was drawn from the permanent civilian employees in the 22 largest Federal executive branch agencies. Of the questionnaires mailed, 16,651 (77 percent) were returned.

Items in the questionnaire typically contained five-point response scales, ranging from "strongly agree" to "strongly disagree." Some items related to general levels of job satisfaction, while others focused on specific aspects of job satisfaction i.e., the nature of the work itself, supervision, various environmental/organizational factors, and behavioral intentions.

The overall level of job satisfaction was moderately high with 68 percent of the Federal work force expressing general satisfaction with their job and 81 percent agreeing that their work is meaningful. Only 18 percent of the respondents reported that they plan to actively look for a new job outside the government, although 31 percent expressed intentions to look for a new job inside the Government.

Although the overall level of satisfaction is moderately high, the results are not uniform across subgroups of Federal employees, as is shown in table 1. In general, the older the worker, the higher the grade level, or the longer the years of service, the higher the level of job satisfaction, i.e the higher the percentage of respondents agreeing with the statement, "In general, I am satisfied with my job."

Even greater than the variations among subgroups, shown in table 2, are the variations among Federal agencies. Overall satisfaction ranges from high levels of satisfaction at the National Aeronautics and Space Administration, to Small Business Administration, and Army (75, 75, 74 percent agreement, respectively) to low levels at the Departments of Housing and Urban Development, Health and Human Services, and Education (56, 55, and 48 percent agreement, respectively). Such variations may reflect variations in the composition (e.g., age, grade or education level) of the work force, agency mission, nature of work performed, and level of funding.

Table 1. Overall Job Satisfaction by Selected Subgroups

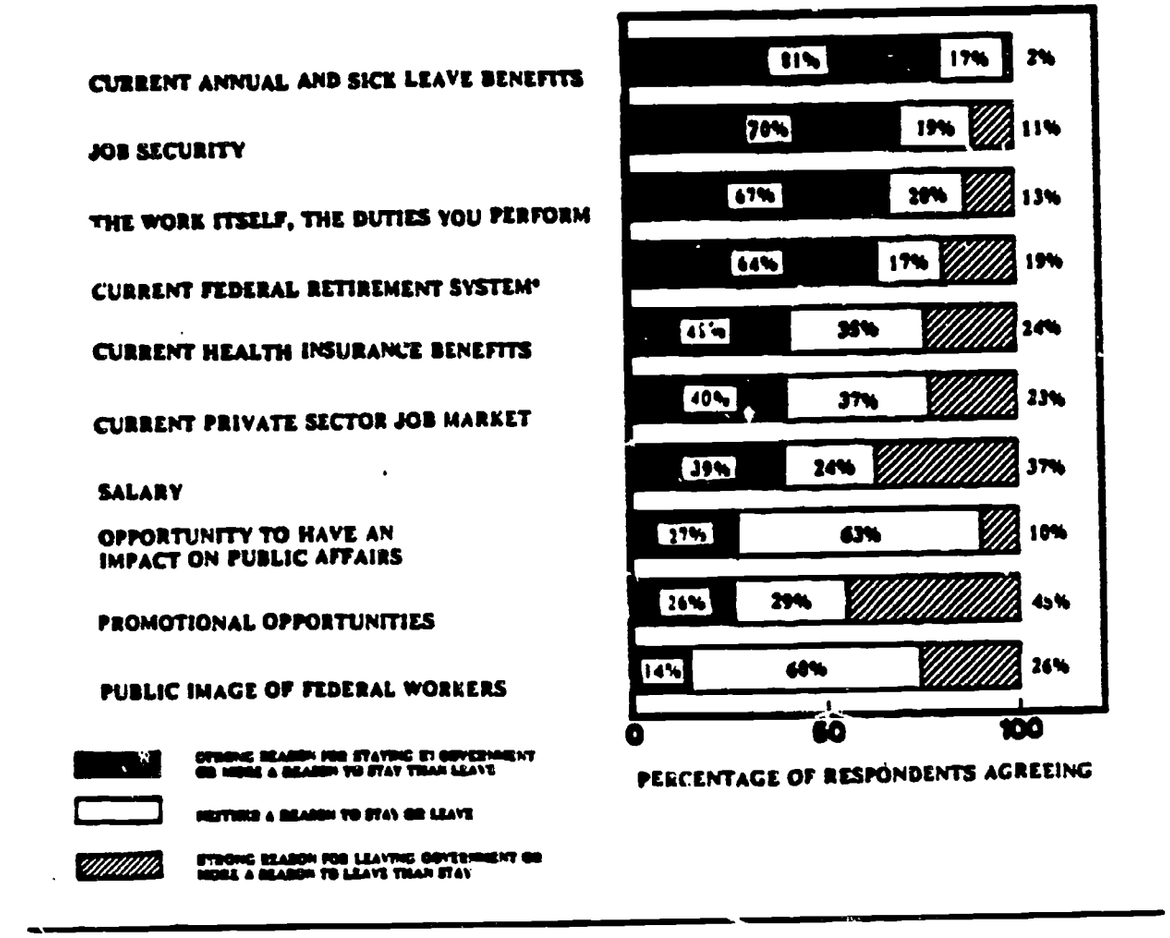
"In general, I am satisfied with my job."

Variable	Subgroup	Percent Agree
AGE	50 years or more	75%
	40 - 49 years	70%
	39 years or less	63%
GRADE	Senior Executive Service	81%
	GS/SM13-15 (Mid-Level)	71%
	GS 9-12	69%
	GS 5-8	67%
	GS 1-4	59%
	Wage Grade (Blue Collar)	72%
LENGTH of SERVICE	20 years or more	77%
	11-20 years	70%
	10 years or less	63%

Figure 1 shows the extent to which various aspects of job satisfaction were cited as reasons for "staying" or "leaving" the Government. Annual and sick leave benefits were cited by 81 percent of the respondents as reasons for staying, followed by job security (70 percent), and the work itself (67 percent). On the other hand, 45 percent cited promotional opportunities (or lack thereof) as a reason to leave, followed by salary (37 percent).

There were also some subgroup differences in the pattern of responses to aspects of job satisfaction. While there were no sex differences in overall levels of job satisfaction or in satisfaction with benefits, fairness of treatment, or supervision, women cited salary, promotional opportunities, job security, and health benefits substantially more frequently than men as reasons for staying in the Federal Government. In addition, top female executives were less satisfied (68 percent versus 81 percent) than their male counterparts, although the opposite was true for women in General Schedule positions 9 through 12 (upper rank-and-file and first-line supervisory positions) were female employees expressed satisfaction 73 percent of the time compared to 62 percent for male employees.

Figure 1. Reasons to Stay or Leave the Government.



Not unexpectedly, older workers considered retirement benefits a more important reason for staying in the government than did younger workers. Less expected was the finding that there were no differences in overall levels of satisfaction between Federal employees working inside versus outside the Washington, DC area or between workers at headquarters versus field locations.

The differences found among groups of employees in their levels of satisfaction with various aspects of worklife lend support to the notion that not all employees are affected the same way by Federal personnel policies and practices. Therefore, efforts to enhance the Federal Government as an employer or to bring about organizational change within Federal agencies should be focused according to these differences. Such efforts are more likely to succeed if they are directed at changing those aspects of work

that are the source of the least satisfaction, and targeting the change to the least satisfied subgroup.

This presentation was based on a report by Jamie J. Carlyle and Paul van Rijn, entitled, Working for the Federal Government: Job Satisfaction and Federal Employees (1988). A copy of the report may be requested from the authors by writing the U.S. Merit Systems Protection Board, 1120 Vermont Avenue NW, Washington DC 20419.

* * * * * * * * * * *

THE RELATIONSHIP BETWEEN RECRUITMENT SOURCE AND EMPLOYEE BEHAVIOR

Michael G. Aamodt & Kimberly Carr
Radford University

Personnel professionals have long been interested in the best ways 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. 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, as it is possible that none of the 50 will perform as well or stay with the organization as long as the two from the employment agency. Thus, the savings obtained in recruitment costs would be nullified by the increased training expenses and reduction in employee performance. While both ideas are important, published research has generally centered on investigating the idea that certain recruitment methods will yield better employees than will other methods.

It was the purpose of this paper to investigate the effectiveness of recruitment source by:

1) Conducting a meta-analysis of related research

2) Collecting new data to investigate if successful employees referred better employees than unsuccessful employees

3) Collecting new data to investigate the relationship between applicant characteristics and applicant utilization of various recruitment methods.

Meta-analytic Review of Previous Research

Five studies were found that investigated the relationship between recruitment source and employee performance and 11 studies were found that investigated the relationship between recruitment source and tenure. Traditional meta-analytic procedures were made difficult due to the small number of available studies, the variety of criteria used, unreported data, and the comparison of different recruitment sources in each study.

So, the first step in this review process was to determine a way of standardizing the data reported in the literature. For example, in one study the tenure data were reported in months employed while in another study the data were reported as a percentage of employees whose tenure was greater than 12 months. To standardize the dates, we took the raw scores for each recruitment method and divided them by the mean for the entire sample. For example, a study reported that applicants answering newspaper ads had an average tenure of 8 months, those who were referred by a friend had an average tenure of 12 months, and those who just walked-in and applied had an average tenure of 10 months. The mean for the study would be 10, and the standard scores for each of the methods, reported as a percentage of the overall study mean, would be 80 for media recruitment, 120 for employee referral, and 100 for direct application.

Once each score in each study was standardized, the scores were averaged across studies to indicate an overall level of relative effectiveness for four recruitment source categories: Employee Referral, Direct Application, Media, and Employment Agencies.

As can be seen in the table below, recruitment source had a significant effect when tenure was the criteria but not when performance was the criteria. More specifically, employee referrals resulted in the highest tenure while media sources resulted in the lowest tenure.

Recruitment Source	Criteria Used	
	Performance	Tenure
Employee Referral	95.60	120.36
Direct Application	102.66	98.89
Media Advertisement	99.03	88.92
Employment Agencies	100.22	91.50

Differential Effects of Employee Referral

As indicated in the table above, employee referrals result in higher tenure than do the other recruitment methods. This finding raises questions about whether all employee referrals are alike. In the only study investigating different types of employee referrals, Hill (1970) compared the performance appraisals received by employees who had been referred by a close friend with the appraisals of employees who had been referred by employees with whom they had only a casual acquaintance. Hill (1970) found no significant effect involving 105 employees in two organizations.

The participants in the current study were 141 former retail and restaurant employees. Each participant was asked to indicate the number of months that he/she worked for the company, who referred them, and the number of months that the referrer had worked at the company at the time he or she made the referral. Referrers who had worked for the company at least 7 months at the time of the referral were designated as "high tenure referrers" while those who had worked less than 7 months were designated as "low tenure referrers." Due to the small number of family members in our sample making referrals, family members were not segmented into high and low tenure groups.

As indicated in the table below, participants referred by high tenure employees and by family members had significantly higher tenure than did participants who were referred by low tenure employees. There was no significant difference between the high tenure and the family member groups. These results indicate that only referrals made by high tenure employees or by family members should be used in recruiting applicants.

Referral Type	n	Tenure
Family Member	17	12.88

Long Tenure Friends	69	11.13
Short Tenure Friends	55	7.69

Such a finding makes a great deal of sense. Research on interpersonal attraction indicates that people are attracted to others who are similar to them on variables such as personality, interests, and attitudes. Thus, an applicant referred by a friend currently employed by the company is likely to be similar to that friend. If the current employee enjoys his/her job, then it is logical to assume that a similar person would as well. Further research is needed to determine if the same pattern will hold for performance measures and if high and low tenure family members differ.

Utilization of Recruitment Source

Two theories have attempted to explain the differential effects of recruitment source on employee performance and tenure. One theory states that informal recruitment sources are superior to formal sources because they provide an applicant with more complete and accurate information than do informal sources. This theory has received empirical support from Quaglieri (1982) and Breaugh and Mann (1984) who found that applicants using informal recruitment sources had more accurate information about the job than did applicants using formal recruitment sources.

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. Research has indicated that applicants who use media sources tend to be male, older, and possess low self esteem (Breaugh & Mann, 1984; Ellis & Taylor, 1983). Applicants who directly apply for a job tend to be female and younger (Swaroff, Barclay, & Bass, 1985; Breaugh & Mann, 1984). Applicants who use employee referrals tend to be younger, while applicants using employment agencies tend to have low self-esteem, and be single (Ellis & Taylor, 1983; Breaugh & Mann, 1984).

To investigate this issue further, 104 students were asked to indicate each job at which they had worked, as well as how they had heard about the job. In addition, the students were given the Employee Personality Inventory (EPI) and asked to indicated their high school grade point average, their sex, and their family income. The five scales of the EPI as well as the responses to the above three questions were correlated with whether or not the subject used any of the four main recruitment strategies in looking for any one of their jobs. Correlational analysis indicated that with the exception of a small correlation between GPA and hearing about the job through a sign posted at the potential place of employment, none of the individual difference variables were related to use of recruitment sources.

* * * * * * * * * *

A COMPARISON OF THE VALIDITIES OF PAPER AND PENCIL MEASURES VERSUS ASSESSMENT CENTERS IN POLICE SELECTION

Joan E. Pynes & H. John Bernardin
Florida Atlantic University
and Donald G. Bergeson, City of Miami Personnel

Two hundred and seventy-five police officer candidates were assessed from 1982 to 1986. The ethnic and gender composition of the candidate sample was as follows: white males = 40; white females = 15; black males = 38; black females = 20; hispanic males = 149; hispanic females = 13.

The data for this investigation came from a one-day assessment program. The assessment center under study was developed in 1981 through a U.S. Department of Justice grant. Three law enforcement agencies were selected to participate in the development of the program. The center exercises and dimensions were based on a job analysis conducted in 1982 (Dade-Miami Criminal Justice, 1982). The job analysis involved interviews and observations of incumbents and supervisors, the administration of a 111 item task-based questionnaire to 1182 police officers, and a factor analysis of the returned questionnaires.

Based on the results of the job analysis, eight "skill clusters" were identified and defined. These clusters were: Directing Others, Interpersonal Skills, Perception, Decision Making, Decisiveness, Adaptability, Oral Communication, and Written Communication. After the skill clusters were identified and defined, a questionnaire was distributed to incumbent police officers who were instructed to rate each skill in order of importance. Perception and decision making were designated as "critical skills". The results of the job analysis were similar to those reported in a review of several multi-jurisdictional job analyses (Bernardin, 1988). Based on the skill areas identified by the job analysis, four assessment exercises were developed.

Formal assessor training programs were conducted after the exercises were developed. Each assessor participated in a three day training program which focused on the assessment exercises and methods for observing and rating performance on the skills (Mendoza & Craig, 1983).

The candidates participated in four assessment exercises in which they were required to assume the position of a police officer. The candidates investigated simulations of a domestic distur-

bance and a homeowner complaint, performed a witness probing, and watched a video simulation of actual or potential crime scenes. The data for each candidate consisted of ratings on eight behavioral dimensions from three assessors, the group consensus ratings for each dimension, and a consensus-derived overall rating which placed each candidate in one of three descriptive categories: 1) less than acceptable, 2) marginal or 3) acceptable.

Performance in the training academy and on the job performance ratings were used as criteria in the validation. The training academy criteria consisted of four written exam scores, scores on firearms proficiency, and two simulations.

Composite measures were derived for the written exams, and the simulations. The last training academy criterion was a composite measure derived by summing the standardized written exam scores and the standardized simulation proficiency scores. The assessment center dimension ratings were significantly correlated ($p<.05$) with the written exam composite and the standardized training academy composite. The overall assessment rating was significantly correlated ($p<.05$) with the written exam composite, the standardized training academy composite, and one of the simulations.

On the job performance was assessed by uncontaminated supervisory performance ratings on 204 police officers. An average of 13 performance ratings were available on each candidate. The uncorrected predictive validity of the assessment center was .20.

* * * * * * * * * * *

A DESCRIPTION OF THE CALIFORNIA PEACE OFFICER STANDARDS AND TRAINING COMMISSION'S COMMAND COLLEGE ASSESSMENT CENTER MODEL AND VALIDATION STUDY

John J. Clancy
Jack Clancy & Associates
Fair Oaks, California

Introduction

The California Peace Officer Standards and Training Commission's (P.O.S.T.) Command College was instituted in 1983 to develop a network of future-oriented law enforcement leaders in the state of California and to prepare those leaders to anticipate,

interpret and confront the issues of law enforcement managers to be the best managers possible and to be capable of successfully addressing the complex management issues which administrators will face in the near future as a result of the quicking space of social and technological changes.

In order to accomplish this, a rigorous two-year educational program was established. This program consists of a curriculum that involves research and forecasting techniques, strategic planning and decision making, transition management, human resource management, public finance, high technology applications and an independent research project.

In order to assess this potential, an assessment process was developed and consists of three phases:

1. The minimum qualifications (MQ's) necessary to be eligible to attend the Command College.

2. The application submitted by candidates which serves as the basis for invitation to participate in the Assessment Center.

3. The Command College Assessment Center

This paper will focus primarily on the Command College Assessment Center.

Definition of Desirable Command College Candidate Attributes

As previously stated, the goal of the P.O.S.T. Command College is to select and train law enforcement managers who have the best potential for meeting future challenges. In order to determine the meaning of "best potential", P.O.S.T. staff reviewed the tremendous amount of research available relative to the characteristics of successful managers. In addition, they talked to many representatives of private industry and major public agencies in order to tap their current thinking on specific traits that could identify outstanding managers in their organizations. The resultant list of attributes are as follows:

1. WRITTEN COMMUNICATION - Effective express written thoughts, ideas, and opinions in clear, concise and accurate language. Anticipates knowledge and needs of reader and prepares complete and well-organized written communications.

2. VERBAL COMMUNICATION - Effectively expresses thoughts, ideas and opinions to individuals and/or groups at all levels. Oral presentations are well organized and tailored to the audience. Handles complex and chal-

lenging questions well. Is articulate and quick to think and respond.

3. INTERPERSONAL RELATIONS - Creates an organized climate resulting in a motivated workforce. Interacts with employees at all levels in the organization. Effective in getting ideas accepted and in guiding a group or an individual toward task accomplish.

4. ENERGY/INITIATIVE - Sets goals and follows through. Actively influences events rather than passively accepting them. Is self-starting. Takes action beyond the minimum required. Originates actions and demonstrates perseverance, personal energy and stamina.

5. JUDGMENT - Demonstrates the capacity to use good sense and wisdom in making reasonable decisions. Recognizes alternatives and assesses the impact on employees, operations and the organization.

6. FLEXIBILITY - Modifies behavioral style and management approach to reach a goal. Is adaptable and deals effectively with diverse views. Has willingness to try different alternatives to find the most successful solution. Considers diverse opinions and approaches in a reasonable manner. Has tolerance for ambiguity.

7. INTEGRITY - Is trustworthy and demonstrates truthfulness in personal and professional activities. Is committed to the ideas and standards of the profession and organization. Acts in accordance with accepted moral values and principles of right and wrong.

8. DECISION MAKING - Develops alternative solutions to problems, evaluates courses of action and makes logical decisions. Establishes priorities and effectively uses available resources to accomplish goals. Takes action or initiates programs where risk of failure is considered element.

9. BUDGET & FISCAL MANAGEMENT - Has Knowledge of operational cost analysis and various budget systems. Has flexibility to adapt existing fiscal resources to support service requirements. Has awareness of competition for and factors affecting fiscal resources. Uses sources of revenue outside of department to expand fiscal resources. Demonstrates ability to clearly communicate budget resources, requirements and limitations within and outside of department.

These attributes (or dimensions) then became the basis for the design of the Command College Assessment Center. It was felt

that applicants for the Command College who possessed most (if not all) of these dimensions would be successful Command College students and graduates. Thus, the aim of the Command College assessment center was to predict successful Command College performance and successful management performance after graduation.

Considerations in the Design of this Assessment Center

In designing the P.O.S.T. Command College Assessment Center, we took the following requirements into consideration:

- We wanted to measure as many of the desirable management dimensions as possible.
- We wanted to use more than one technique to measure each dimension.
- We wanted a range of measurement techniques in order to be able to conduct research to identify the kind of techniques which were giving us the most accurate information.
- We wanted the techniques to be as job related and relevant as possible.
- We wanted to evaluate up to 50 applicants in one day.
- We wanted to be able to identify those applicants who would be accepted into the Command College on the same day as the assessment center.

P.O.S.T Command College Assessment Center Model

The P.O.S.T Command College Assessment Center is a one-day evaluation process consisting of the following measurement techniques:

- A Leaderless Group Discussion
- Two individual Interviews: Past Experiences and Life Goals
- Written Tests: an essay writing exercise, a test of critical thinking and a personality test

The assessment center process was designed to evaluation up to 48 candidates in four 1 and 1/2 hour sessions. The 48 candidates are divided into four groups of 12. Each group of 12 receives a different order of presentation of the measurement techniques. For example one group would be evaluated in the following manner:

8:30 - 10:30	Leaderless Group (two groups of six candidates each)
10:30 - 12:30	Critical Thinking Test
12:30 - 1:30	Lunch
1:30 - 3:30	Past Experience Interview & Personality Inventory
3:30 - 5:30	Life Goals Interview & Essay Writing Exercise

Four assessors are needed to rate the Leaderless Group Discussion performances (two assessors for each group of six) and eight assessors are required to conduct the 24 interviews (which last approximately 20 minutes each). At the end of the eight-hour process, each of the 48 candidates has been given seven independent evaluations - (1) the Leaderless Group Discussion Rater #1; (2) the Leaderless Group Discussion Rater #2; (3) the Past Experience Interviewer; (4) the Life Goals Interviewer; (5) the graders of the Essay Writing Exercise; (6) the Psychologist's review of the critical thinking test results; and (7) the Psychologist's review of the personality inventory results.

P.O.S.T Command College Assessment Center Decisions Making Process

For each candidate, a tremendous amount of information has to be combined into one final decision - should the individual be admitted into the Command College. The decision making approach we selected requires that each evaluator decide whether he/she thinks the candidate should be accepted or rejected based sclely on the individual evaluator's data. Using this method, each of the six evaluators gets a "Vote" - the Leaderless Group Discussion Rater #1, the Leaderless Group Discussion Rater #2, the Past History Interviewer, the Life Goals Interviewer, the Essay Writing Grader, and the Psychologist (based ont he results of the critical thinking test and the personality inventory).

Based upon the pattern of "YES" and "NO" votes, candidates will fall into one of three categories - ACCEPT, REJECT, and DISCUSS. These categories are not designed to produce completely automative decisions relative to an individual's candidacy for the Command College. Rather they are preliminary recommendations which are presented to the assessment center evaluators. The final decision is made in an assessors' consensus session held immediately after all the evaluations have been made and the data summarized. Here, the names of the candidates in the ACCEPT and REJECT categories are presented to the evaluators and finalized unless a specific objection is raised. Each candidate in the DISCUSS category is then discussed in detail and assigned to either the ACCEPT or REJECT category.

Command College Validity Study

We are now in the process of conducting on-going validity research into the effectiveness of the Command College assessment process. This study consists of the following components:

- The data that is gathered in the Command College application process;
- The assessment center measures;
- Criterion measures used to evaluate student success in the Command College program and back on the job.

The data gathered in this validity research will help evaluate the components of the assessment process which have been established to select students into the Command College and also the content of the Command College curriculum. The result of this evaluation will be the kind of data which are needed to make a step-by-step alteration and improvement in the selection and training process. The ultimate goal is a selection and training process which chooses the best candidates to enter the Command College and gives them the kind of preparation they need to become effective future law enforcement leaders.

* * * * * * * * * * *

THE ASSESSMENT CENTER: REDUCING INTERASSESSOR INFLUENCE

Phillip E. Lowry
University of Nevada, Las Vegas

There is little reported research on the consequences of variations in assessment center procedures. Cohen (1978) has suggested that the consensus discussion is the most central aspect of assessment center technology. Silverman, et.al. (1986) pointed out that an important aspect of the assessment center is the way evaluations of participants are made by the assessors. Sackett and Wilson (1982) suggest that the consensus judgment process includes the opportunity for some assessors to exert more influence on the outcome than others.

The purpose of this paper is to report on the use of consensus discussion procedures designed to reduce the influence an assessor may have on others.

Sackett and Wilson have suggested (1982) "differences in (assessor) influence are a phenomenon worthy of further consideration."

They based their finding on the observed differences in influence in two assessment centers.

They operationalized the assessor's influence on the consensus decision as the frequency with which an assessor changed a rating during the consensus discussion. Having an assessor's rating adopted by the group was evidence of being influential. Hence the smaller the relative number of scoring changes, the greater the influence of the assessor.

While there may be other factors that would explain why assessors change their scores, the influence factor suggested by Sackett and Wilson (1982 was accepted as the basic premise for this research.

This paper presents finds about interassessor influence observed in four assessment centers. Each center included consensus discussions that were conducted following a procedure designed specifically to reduce interassessor influence.
Delbecq, Van de Ven, and Gustafson (1975) developed a procedure to minimize the domination of a group by one or more individuals. Their process, Nominal Group Technique (NGT), was developed specifically to deal with decision making by small groups. Such group sessions require pooling of judgments; and such groups can be dominated (whether for good or bad) by one or more individuals.

The consensus procedure described in this paper was based on the Nominal Group Technique. The basic research questions was whether this procedure would reduce interassor influence.

Method

Data for this study were collected during four centers conducted for local governments. Eighteen individuals were rated on five dimensions by seventeen assessors. THe assessors in the selection centers were generally homogenous with respect to position, training, and experience. The assessors in the career development centers were not.

Two scores were developed by the assessors for each participant on each dimension; a pre-pooling score on each dimension (the raw arithmetic score before any discussion),a nd the score developed after the consensus discussion.

Four different types of simulation exercises were used in each center: a written analysis of three critical events, a written/oral analysis of a problem, a role playing exercise involving a personnel problem, and a leaderless group discussion.

The procedures used to observe and evaluate the participants were the same in each assessment center. The assessors were senior

level managers who were trained in the assessment center process. In the selection center, each assessor had the opportunity to observe all the participants in each exercise. In the career development centers not all assessors were able to observe each participant in each exercise; however, at least two assessors evaluated each participant during each simulation exercise.

At the conclusion of all the exercises the assessors prepared a summary of the IMPORTANT behaviors they had observed throughout the exercise. They classified these behaviors under the appropriate performance dimensions and recorded and "initial", pre-consensus score. They were told that these scores would be subject to change during the consensus discussions. These pre-consensus scores, like all other scores were never attributed to an assessor.

On the following day the assessors participated in a series of discussions to arrive at a consensus on the scores for each performance dimension. The director coordinated the discussions, but did not participate in them nor provided any input into the scoring process. The discussion followed the general procedure for the Nominal Group Technique described by Delbecq, Van de Ven, and Gustafson (1975).

For each participant, each assessor, in turn, discussed the behaviors they observed that related to each of the performance dimensions. After the discussion on each performance dimension, the assessors were told to give the participant a score on the dimension based on not only their own observations, but on the observations reported by the other assessors. This scores was given to the director of a slip of paper and never attributed to the assessor. Nor discussion of scores was permitted at any time. It was assumed that by not attributing a specific score to an assessor, the other assessors would be more likely to exercise independent judgment.

If the scores were within one rating scale (a continuous scale of 1 - 5 was used), consensus was obtained. This score was recorded as the assessor's post-consensus score. (See Sackett and Wilson, 1982 for a precedent for using one rating scale or less as reflecting consensus).

If there was more than one scale difference, the assessors were asked to conduct another iteration of the discussion of behaviors and to elaborate on these behaviors to ensure non were overlooked. They then resubmitted the scored. his was the final score even if there was more than a one point difference in the range.

The number of changes in scores from the pre-consensus score to the post-consensus score on each dimension for each assessor was calculated.

Results

A one-way univariate analysis of variance (ANOVA) indicated that the difference in scores across assessors was not significant in any of the assessment centers. The results of the ANOVA are displayed in Table 1.

Table 1
Changes in Scores by Assessors

Assessment Center	Mean number of changes	F	Significance
Career Development 1	3.72	1.474	0.32
Career Development 2	4.69	0.231	0.87
Selection 1	4.25	1.176	0.36
Selection 2	1.50	2.296	0.16

Total number of cases for analysis = 240

Discussion

Sackett and Wilson (1982) reported on two assessment centers. One center was a low level management center; the other was for high level management. They found a significant difference in the number of rating changes in the high level center, and no significant difference in the low level center. The assessment centers used in this research were for high level management positions.

Sacket and Wilson (1982) did not report on the precise consensus procedures used in their high level assessment center. They did report on the procedures used in the low level center. It is assumed that the same procedures were used in both. These procedures differed from the consensus procedures in this research primarily in that the assessors revealed their ratings on each dimension.

One of the salient features of the consensus procedures detailed here is the confidentiality of the ratings. At no time were the assessors permitted to divulge their scores. They could and did attempt to describe behaviors; they were not allowed to disclose their evaluation of these behaviors.

Conclusions

The purpose of this paper was to report on the results of using a consensus procedure that was designed to reduce interassessor influence. No significant interassessor influence was found in the four assessment centers. The consensus procedure used may have contributed to these results. However, there is insuffi-

cient evidence to suggest that the procedure alone reduced the influence. There may have been other factors, including the skill and training of the assessors, the behavioral characteristics of the assessors, the quality of the centers, and other similar factors.

The reported on consensus procedure is based on a sound and proven technique, and it does appear to be a reasonable way to conduct the pooling process. Additional research is required to validate the proposition that this consensus procedure can invariably reduce interassessor influence. Practitioners may wish to consider using this procedure despite the lack of complete validation.

References

Cohen, S.L. Standardization of assessment center technology: Some critical concerns. Journal of Assessment Center Technology, 1978, 1, 1-10.

Delbecq, A.L., Van de Ven, A., and Gustafson, D.H. Group Techniques for Program Planning. Glenview, Ill.: Scott, Foresman, and Company, 1975.

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

Silverman, W.H., Dalessio, A., Woods, S.B., Johnson, Jr., R.L. Influence of Assessment Center Methods on Assessors' Ratings. Personnel Psychology, 39, 1986, 565-578.

Sackett, P.R., and Wilson, M.A., Factors Affecting the Consensus Judgment Process in Managerial Assessment Centers. Journal of Applied Psychology, 1982, 67, 10-17.

* * * * * * * * * *

VALIDATION OF PHYSICAL PERFORMANCE TESTS

Carolyn E. Crump & Deborah L. Gebhardt
Advanced Research Resources Organization
A Group of University Research Corporation
Chevy Chase, Maryland

Validating selection tests for entry into physically demanding jobs requires a detailed job analysis, an understanding of the working environment, and knowledge of testing human capabilities. This paper will focus on the latter two requirements and describe how an understanding of the working environment and the application of physiological principles contribute to developing, validating, and transporting physical performance tests.

Understanding the working environment regarding the implementation of entry-level tests helps guide the development of the physical performance tests and criterion measures. The working environment takes into account the ergonomic parameters such as heights, weights, forces, etc. coupled with issues related to frequency and time spent, that are experienced by the employee. Ergonomic factors must be identified in the work environment to ensure that the tests and criterion measure(s) adequately reflects the physiological demands of the job. Other factors include the ability of the employer to assign the applicant to a variety of entry-level positions and the financial and personnel resources available for implementation of the validated tests.

Development of Physical Performance Tests

Two types of tests have been used to measure physical abilities: (1) basic ability tests and (2) job sample or simulation tests. Basic ability tests are developed to measure the abilities required to perform adequately in a job. Job sample or simulation tests include components of the job being studied (e.g., climb a ladder) and might require an applicant to use equipment used on the job. Simulations are typically limited to a specific job. Four issues are considered in deciding to use either basic ability or job sample tests: validity, adverse impact, safety, and practicality.

Validity. Many of the studies on physical ability selection testing have used basic ability tests. Evidence has now accumulated that basic ability tests have significant criterion-related validity for a variety of physically demanding jobs (e.g., Arnold, et al., 1982; Braithwaite & Markos, 1980; Chaffin, Herrin, Keyserling, & Foulke, 1977; Crump et al., 1985; Gebhardt et al., 1983; Gebhardt, Crump, & Schemmer, 1985; Gebhardt, Schemmer, & Crump, 1985; Gebhardt & Weldon, 1982; Reilly et al., 1979). Although few studies have compared the relative validity of basic ability and job sample tests, several have found that the use of basic ability tests resulted in a higher or similar multiple correlation (R) with the job performance measure (e.g., supervisor ratings) (Crump et al., 1985; Hogan, Jennings, Ogden, & Fleishman, 1980, Hogan Ogden, & Fleishman, 1979; Wunder, 1981).

Adverse Impact. Physiological research and test validation research in the area of physical performance has shown that there are significant gender

differences in both basic physical ability tests and job sample tests. In the studies that incorporated both basic ability and job sample tests, the magnitude of the gender differences were similar for both the job samples and the basic ability tests with the men generally scoring higher than the women. Differential prediction analyses indicated that the basic ability tests were fair to men and women (i.e., no slope difference) and minorities.

Safety. Safety is of particular concern when one considers the wide range of applicants (e.g., age) that may be tested as a result of the removal of laws and statutes limiting the applicant pool (e.g., Age and Discrimination in Employment Act, Rehabilitation Act of 1972). Basic ability tests are easier to administer and can be monitored in relation to the applicant's safe response to the testing protocol.

Practicality. Using the basic ability approach, the number of tests is limited to the number of abilities required by the jobs and is independent of the number of physically demanding tasks in the job.

Criterion Measure

The job performance measure used for validating physical performance tests must meet several criteria. First, the criterion measure must be relevant and important to performance of the physical aspects of the job. Therefore it must reflect the physiological parameters of task and job performance. Second, the measures must be reliable and not be contaminated by non-physical job performance dimensions. Third, the criterion measure must discriminate between employees who are adequately performing the physical aspects of the job and those who are not. Finally, the criterion measure must be practical and safe, and not interfere with daily work or production. Several types of criterion measures used to validate physical performance tests for manual materials handling, manufacturing, and public safety jobs are highlighted. Three types are supervisor and/or peer ratings of (1) job tasks, (2) a combination of physical abilities and job tasks, or (3) physical abilities and the fourth type described is a work sample.

Ratings of job tasks. To validate the physical performance tests for the selection of paramedics, peer ratings of critical job tasks were employed (Gebhardt & Crump, 1984). Two steps were taken to select critical tasks which were representative of the relevant physical abilities for use in the criterion measure. The ten highest rated critical tasks for each physical ability were reviewed in relation to their mean frequency rating. For each task selected, six behavior descriptions of task performance varying in degree of difficulty and outlining superior to inadequate levels of performance were developed. Each level of the behavioral descriptions contained specific information obtained in the job analysis related to weight, body position, distance, time, etc. and incorporated the physiological demands. Adequate or acceptable performance was determined from the job analysis results and was defined as level four on the one to six scale.

The reliability of the peer ratings was determined using a model that evaluated the reliability of multiple raters for a single paramedic (Shrout & Fleiss, 1979). The interrater reliability coefficients for two raters ranged from .49 to .66 for the seven tasks. The final criterion measure consisted of a

unit weighted sum of the task ratings and an overall physical job performance rating and resulted in a multiple correlation of .61 with three physical performance tests (i.e., dynamic lift, modified stair climb, arm lift).

Work sample and ability rating criteria. A second study in the tire manufacturing industry involved the use of two criterion measures, a work sample and supervisor ratings (Crump, Gebhardt, Guerette, & Wertheimer, 1985). The objective of the research was to develop and validate a single test battery that could be used to select individuals for seven different jobs. Therefore, the criterion measure developed had to be applicable to all seven jobs. The results of the job analysis indicated that there were five physical abilities that were common to the seven jobs. The job performance measure was applicable to all seven jobs.

The supervisor rating criterion measure consisted of ratings of the physical abilities with examples of critical, frequent job tasks listed beneath the ability definition. For each job, different tasks were listed below the ability definition. A seven-point scale related to basic job requirements was selected because supervisors had experience with evaluating workers in relation to production standards and requirements. The interrater reliability estimates for the five ability ratings ranged from .63 to .84 for two raters.

Since the supervisor ratings for the five abilities were provided in relation to a specific job and not across all jobs, the ratings were rescaled to reflect the different mean levels in each job for the specific physical abilities obtained in the job analysis. This rescaling ensured that ratings given by supervisors in one job would be equivalent in magnitude to the ratings given by supervisors for other jobs.

For the work sample the highest rated top one third of the critical tasks on each physical ability were reviewed for each job. These tasks were clustered into four movement categories: lift, push, pull, and carry. Review of the job analysis results indicated that the ergonomic parameters such as weight of materials, height lifted to, and plane of movement were similar across jobs. Based on the movement categories, physiological demands, and ergonomic data, three work sample criterion measures were designed that consisted of sequences of activities that were related to the frequent and important tasks and found in all seven jobs.

The scoring system developed for each task allowed individuals who were unable to perform all segments of a task to complete the work sample. The reliability of the work sample tasks was determined with a test-retest approach. The test-retest correlations for the sidewall/push-pull, tire sort, and bale lift ranged from .68 to .80. The split halves correlations (N=245) for internal consistency ranged from .75 to .87. The work samples were standardized and summed for the validity analysis and the rescaled supervisor ratings were summed. The correlation of the two measures was.41.
The multiple regression analysis for the work sample resulted in a correlation of .82 with three physical performance tests (i.e., arm endurance, arm lift, arm power). When the supervisor ratings were used in the multiple regression analysis, the multiple correlation was .47 and yielded the same predictor tests as the work sample.

Transportability of Physical Performance Tests

Transportability allows for the use of a selection instrument validated for a job in one organization to be used by a second organization if the jobs in each organization are similar. A transportability analysis involves a systematic comparison of the job in the first organization to the job in the second organization. A transportability approach is an efficient and cost effective method to determine whether the same physical performance tests validated for one organization can be used for selection into a similar job for a second organization.

The transportability procedure is outlined in the Uniform Guidelines Rules and Regulations (1978, p. 38299). This Federal document indicates that specified criteria are required to transport tests. The criteria are as follows: (1) criterion-related validity evidence must be present; (2) validity evidence must show that the selection procedure is valid; (3) incumbents in the "new job" must perform substantially the same job tasks in the original job; and (4) evidence must be provided which indicates that the tests are fair to minorities (e.g., ethnic, gender, race).

Determination of job similarity. This procedure consists of determining the percent overlap between the original job and the new job. This is based on an examination between the common and unique critical job tasks. If the results of this analysis yield an 80% or greater overlap the jobs are considered similar and the same physical tests that were validated for one organization may be used for selection by a second organization. If there is not an 80% overlap in the job similarity analysis, an abilities approach may be employed. This approach is recommended in the Joint Standards for Educational and Psychological Testing (1985). This approach compares the abilities required to perform one job with those required to perform another job, even if the tasks are not similar. The physiological aspects are incorporated in this approach.

Conclusions

Based on a thorough understanding of the job, work environment, and physiological principles, valid and reliable physical performance tests can be developed and used for selection into a variety of jobs. The criterion measures used in criterion-related validation studies must be based on the critical job tasks and may involve several different formats (e.g., supervisor or peer ratings of tasks or ability, work sample tasks). Basic ability tests have been found to be fair, valid, safe, and practical for selecting applicants for a variety of physically demanding jobs. Further, the tests validated for a job in one organization may be transported to another organization if similar tasks are performed or abilities required and if criterion-related evidence exists that indicated the tests are fair to all protected groups.

References available upon request

* * * * * * * * * * *

Is a Uniform Guideline for Fitness Tests Possible?

Vernon R. Padgett
and Gene Carmean

Med-Tox Associates, Inc.
Tustin, California

Congress recently passed major legislation impacting on employment policy. As of January 1987, the mandatory retirement age of 70 no longer exists. Police and firefighters, however, are temporarily exempted. Arbitrarily-selected entry ages as low as 31 years still apply in some jurisdictions, and can continue until the end of 1993. The Equal Employment Opportunity Commission (EEOC) and the Department of Labor have been mandated to investigate the validity of mental and physical fitness tests, which could serve as substitute for age in retirement decisions.

Mandatory Retirement is Unfair

The origins of mandatory retirement laws can be traced back to Otto von Bismarck's selection of age 65 for payment of retirement benefits under the German social security system. Bismarck selected that age over a century ago, when life expectancy was half what it is today. Had Chancellor Bismarck picked another age, that age would be considered our "normal" retirement age. Many agencies have no retirement age. In the Fire and Emergency Services Department of Hobbs, New Mexico, for example, no age discrimination currently exists. Some firefighters are 60 and 61. The police training officer is 58, and runs 4 miles every day.

Mandatory retirement ages are particularly unfair today. Age is no longer as relevant a criteria for employment as in the past. Today, Americans are more aware of the benefits of physical fitness than ever before. One reason for this change has been increased public awareness that medical science does not have all answers to increasing life expectancy. Americans have shouldered a greater responsibility for health maintenance. Evidence for this claim is found in a number of areas: Increasing concern with diet, legislation against smoking, a decrease in sales of cigarettes, and most markedly, by the physical fitness revolution.

A Revolution in Attitudes Towards Fitness

There are more older Americans than ever before (1), and they are more aware of health and fitness issues. These attitudinal and demographic shifts have unmistakeable implications for the workplace as older workers resist forced retirement. Dramatic evidence for the personal awareness of health is seen in the decreased heart disease each year since 1965, partly attributable to changes in lifestyle (2). Exercise benefits psychological health as well as physical health. Recent research by experimental psychologists indicates that exercise improves mood (3), reduces depression (4), and increases energy while decreasing tension (5). Another change in American's attitudes towards fitness is reflected in the promotion of health at the workplace. Some writers claim that increased physical fitness among American workers would save billions of

dollars in reduced sick time and improved productivity.

A Fitness Decline in America

Even though Americans are now more aware of the benefits of physical fitness, and even though many are fitter than ever before, the general level of physical fitness is poor. For example, 90 percent of females over the age of 16 cannot do more than two pullups (6). The President's Council on Physical Fitness has reported that American youth have made no improvement in physical fitness since 1975 and that American youth scores very poorly in all areas of fitness including cardiorespiratory, strength, agility, and flexibility measures (6,7). This decline in overall fitness has had an adverse impact on employers searching for workers for jobs which require physical fitness and ability. For example, the California Highway Patrol found that CHP officers were in worse physical condition than the average state prison inmate and a "disturbing number" were at "unacceptably high risk of heart attack" (8).

Age and Physical Performance

Many researchers argue that chronological age is not a particularly meaningful variable when assessing physical performance, especially job performance (e.g., 9). Workers vary in their ability to do the job, and a fairer measure of job-related ability than chronological age currently seems appropriate (e.g., 10).

Differentiating among workers by fitness appears fairer than making the decision by age, particularly when the job requires high physical fitness. The fairness of shifting from age to fitness hinges, first, on the assumption that fitness can be measured accurately, and second, that fitness is a better predictor of job performance than age. Congress therefore requires clear answers to several questions. These include: What constitutes job-related fitness? Can job-related fitness be measured accurately? And is it more important that police and firefighters be fit or be young?

Are Fitness Tests Valid?

The purpose of the Congressionally-mandated research is to investigate whether fitness tests can measure the abilities required by police and firefighters. This program will proceed with six steps: 1) Identification of critical tasks performed by police, firefighters, and corrections officers; 2) Analysis of these tasks, which will determine the physical, medical and psychological variables that are critical in task performance; 3) Assessment of the existence of valid and reliabile measures of these variables or their potential for development; 4) A survey indicating the extent to which agencies are currently using such measures for selection and retention; 5) An assessment of the extent to which public safety agencies are using accepted validation procedures in developing such tests; and finally, 6) a cost/benefit evaluation of the use of such fitness tests. Several useful methodologies exist by which these steps may be achieved. These are summarized below:

Comprehensive Research Review with Meta-analysis

One approach is to review areas concerned with aging, fitness, and job performance. Meta-analysis is an innovative, relatively recent method for integrating large bodies of research (11). The basic idea of meta-analysis is to apply the attitude of data analysis to quantitative summaries of individual

studies. Individual studies are aggregated, and weighted according to their importance, with importance judged on such features as sample size, statistical significance, methodological rigor, and size of effect. An example of the ability of meta-analysis to bring clarity to muddled research findings in the job performance area was offered by Waldman and Avolio (12). These researchers addressed another apparent conflict: Some studies on job performance had shown that older workers performed more poorly than younger workers; other studies claimed that older workers performed better. They found that workers are more productive as they get older, when the measurement is objective (like productivity measures), but performance decreases when it is measured subjectively, as with supervisor ratings.

Meta-analyses planned for the fitness study include a review of treadmill testing studies, with Age added as a variable, to regress sensitivity and specificity on age. This may indicate whether these variables change as function of age, and thus whether the validity of the treadmill test changes as a function of participant age. Another useful application of meta-analysis involves reviewing test validation studies to determine the optimal interval for administering fitness tests. The dependent measure would be a composite fitness measure and the predictor variable would be time between testing (Test Interval).

Comprehensive National Survey of Common Practices

A second approach toward answering whether fitness tests are valid involves a survey records held by Police and Fire Departments.

Reanalysis of Existing Data Sets

The third approach to this large-scale effort to comprehensively study the nation's fitness testing is concerned with research data already collected. In this phase of the overall project, data from existing physical ability testing programs will be reanalyzed with Age introducted as new variable. By so doing, the role of age may be assessed without the expense of designing and carrying out original data collection.

Experimental Investigation on Determinants of Fitness

The experimental approach involves gathering medical, physiological, and physical fitness scores on a variety of physical abilities tests (aerobic capacity, dynamic upper body strength, etc). Performance will be measured on job task simulations. Examining the magnitude of statistical association between physical abilities and job task performance would allow an evaluation of the relationship between fitness and job performance. Similarly, the magnitude of association between age and job performance could be assessed.

Another use of the experimental method involves validating visual acuity standards. By experimentally controlling visual acuity (through "decorrective" lenses), a variety of levels of visual acuity can be subjected to empirical test in critical job performance scenarios. A level of minimally acceptable uncorrected visual acuity could be specified as a result of such tests for different classes of public safety officer.

Is Fitness a Fair Basis for Discrimination?

Regardless of the outcome of the study, discrimination will still take place

in decisions to hire and retain. No greater number of police, firefighters, and corrections officers will be hired than before. An equal number of applicants will be disappointed. The difference will be a change in the dimension on which hiring decisions are made. The question remains: Will the new criterion be fairer? Is it fairer to turn away an unfit 21 year old than to turn away a fit 65-year old?

References

1. Ryan AJ. Corporate wellness programs and health benefits coverage. Fitness in Business, 2:121, 1988.

2. Paffenbarger R, Hyde R, Wing A, Steinmetz C. A natural history of athleticism and cardiovascular health. Journal of the American Medical Association, 252:491-495, 1984.

3. McCann I, Holmes D. Influence of aerobic exercise on depression. Journal of Personality and Social Psychology, 46:1142-1147, 1984.

4. Folkins C, Sime W. Physical fitness training and mental health. American Psychologist, 36:373-389, 1981.

5. Thayer R. Energy, tiredness, and tension effects of a sugar snack vs. moderate exercise. Journal of Personality and Social Psychology, 52:119-125, 1987.

6. President's Council on Physical Fitness and Sports 1985. National School Population Fitness Survey, p. 69.

7. National Center for Health Statistics: Health, United States, 1986. DHHS Pub. No. (PHS) 87-1232. Public Health Service. Washington. U. S. Government Printing Office, Dec. 1986.

8. Presentation to the US Senate Public Employee's Retirement Committee, J Voss, Asst Chief, Personnel and Training, California Highway Patrol, 11 Dec 1981.

9. Sharkey B. Functional versus chronological age. Medicine and Science in Sports, 19:174, 1987.

10. Cady L, Phillip C, Karwasky R. Program for increasing health and physical fitness of firefighters. Journal of Occupational Medicine, 27:110-114, 1985.

11. Glass GV, McGaw B, Smith ML. Meta-analysis in Social Research. Sage, Beverly Hills, 1981.

12. Waldman DA, Avolio BJ. A meta-analysis of age differences in job performance. Journal of Applied Psychology, 71:33-38, 1987.

13. Carmean G. Police management considerations of physical capacity screening. The Police Chief, 42-44, January 1984.

Use of Departmental Ratings of Promotability

in

Promotional Examinations

Carol Morris, Senior Personnel Analyst II

City of Los Angeles

In 1975 the Personnel Department of the City of Los Angeles started using Department Ratings of Promotability (DROP) in a small number of high level Civil Service examinations. Since 1982, they've been used as weighted parts of certain examinations. This paper describes the City's efforts to gain one union's acceptance of the process.

Background

In 1983 the Personnel Department, at the behest of the Department of Water and Power, made the Departmental Rating of Promotability a weighted part of the Civil Service promotional examination for Senior Power Engineer. When results were published, an unfair (employee relations practice) was filed by Engineers and Architects Association on behalf of some of the candidates in the examination who had done poorly on the DROP portion. They argued that they had never been told that their performance was marginal or unsatisfactory.

In November, 1984 the issue was heard by the Civil Service Commission which reaffirmed the use of DROP, but with maximum weights of 40% in a two part exam and 25% in a three part exam.

In a hearing before the Employee Relations Board, the plaintiffs requested that they be allowed to review comments made by raters and all of their working papers showing how they arrived at their conclusions. The City's Attorneys argued successfully that such papers should remain confidential to protect the identity of persons making ratings. In a prior case, the Board had already ruled that selection was properly within the jurisdiction of the Civil Service Commission and not subject to employee relations process. The union did not consider the issue resolved, however.

In 1985, Personnel Department staff met with representatives of Engineers and Architects to try to resolve lingering concerns about the DROP. The union's main concern was the use of the DROP in the Department of Water and Power. Their position was that such a test gave management too much latitude to ensure objectivity. They asked that they be allowed to review all of the documents and information related to the DROP so that they could be assured that no manipulation had taken place. Staff argued that such documents were confidential. The union countered that confidentiality shouldn't be an issue because candidates had signed releases allowing publication of the information. They further argued that the DROP wasn't an objective part of the examination process, nor did it contain questions to be answered by future examinees, so disclosure didn't put the examination at risk.

Staff was still concerned, however, about protecting the identity of the raters and the confidentiality of their comments. Staff also feared that candidate acquisition of specific test instruments prior to test administration would compromise the integrity and impartiality of future examinations, in that those candidates who had received information through competition in the previous examination would have an unfair advantage in the new examination.

Use of the Related Achievement Record with a Departmental Rating of Promotability

The Related Achievements Record (RAR) was originally intended as a selection device. It had been used in a few examinations both as a separate test and as a supplement to the candidate's application.

In terms of format, the RAR includes four to six dimensions (factors) which are determined by the examination analyst during the study of the class to be tested. The study may include a job analysis or a conference with incumbents and supervisors to discuss the class. The initial step in the use of both the RAR and the DROP is to determine the tasks involved and the elements critical to successful performance in the job. If for example, problem solving skills are crucial to effective performance in a job, a narrative description is developed to define the category and the candidate is required to describe two accomplishments in that category which would clearly demonstrate his or her skill.

As in the RAR, the Departmental Rating of Promotability includes a problem solving category with the same definition as that of the Related Achievements Record. Unlike the RAR, the DROP includes rating scales such as satisfactory, unsatisfactory, outstanding, etc. Each scale further delineates the kinds of behaviors typical of a given performance level. The rater, then, is asked to make an assessment based upon independent observation of the candidate's performance, as well as the candidate's own description of his/her achievement in each of the areas.

The intent of the DROP is to assess a candidate's potential to assume higher level responsibilities, much the same as interviewers do in an interview situation. While they do not include an assessment of past performance, they are concerned with those aspects of an employee's performance which indicate probable ability to perform at a higher level.

The DROP is like the interview in that a structured rating sheet is used which includes factors identified as critical to job success, and a description of behaviors sought in each of the factors. Unlike the interview process, in which judgments are made on information presented without reference to performance, the DROP is based entirely on performance.

For example, oral communication skill is required in almost every job. More or less of the skill may be dictated by the level of the job. Let's say that for the position of Senior Power Engineer the ability to communicate effectively is key to a person's success in the job because Senior Engineers are routinely required to represent the Department of Water and Power in meetings with heads of other agencies, present information before legislative bodies, communicate with subordinates, etc. Thus, that factor on a DROP would look like this:

Oral Communication Skill

The ability to convey ideas clearly and concisely; explain simple and complex information with equal ease; able to focus on main point of question and not ramble off the subject.

Superior - Based on past performance, if promoted candidate will consistently demonstrate superior ability to convey ideas clearly . . .

Satisfactory - If promoted candidate will usually demonstrate average ability to . . .

Unsatisfactory - If promoted candidate will demonstrate limited ability to . . .

After reviewing the candidate's accomplishments as described in the RAR, the raters express a judgment about him/her in numerical terms as to his/her probability of success at a higher level. The candidate receives the average of the two raters' scores in each category, and an overall score reflective of individual averages.

The advantages of using an RAR with a DROP are to allow the candidate to provide input into his/her evaluation process as well as to give the supervisor another perspective of the candidate's job performance. Based on this modification of the process, the Engineers and architects association withdrew its unfair.

Candidate concerns about DROP

Generally candidates believe that their job performance should be evaluated and considered in the exam process, but some have expressed concern about potential abuses by supervisor and managers and many do not fully understand the process or its purpose.

Analyst Responsibilities

1. Briefs raters
   - explains purpose of DROP
   - reviews rating factors and scales
   - stresses fundamental differences between DROP and performance appraisal

- stresses confidentiality
- urges independent grading and need for raters to support their grades with comments.

2. Collects rating sheets and reviews for adherence to instruction, appropriateness of comments and tries to resolve discrepancies.

Final Review Period

Candidates receive overall score and analyst paraphrases rating factors and any comments. Candidates are allowed to protest fraud, prejudice, clerical error.

* * * * * * * * * *

PERSONALITY TESTING

Donna L. Denning
Personnel Reseach Psychologist

City of Los Angeles

Cognitive tests predict job performance quite well. This holds true for a wide variety of cognitive tests, both general mental ability to learn the job and job knowledge needed to do the job, across a wide variety of jobs. Use of these tests for personnel selection helps to ensure that employees have the ability to learn/do the job for which they are hired; they help to answer the question: "Can this person do the job effectively?

In discussions with supervisors about variations in employee job performance, a counterpart concern usually surfaces: " Will the person do the job effectively?" Certain behaviors facilitate achievement of a high quality and quantity of work: Reliability (showing up for work), punctuality (showing up on time), initiative (doing routine tasks without being told; reporting problems; suggesting improvements), and teamwork (helping out co-workers when there's a need) are examples of these behaviors. Exhibition of these behaviors is relatively independent of mental ability. They are more likely a reflection of certain personal characteristics which include interests, values, temperament, personal history (biographical information), and dimensions of personality. These personal characteristics comprise the area of measurement known as noncognitive testing.

Noncognitive tests, often loosely referred to as "personality tests", are often misunderstood as a personnel selection device. One factor contributing to this misunderstanding is the tendency to confuse clinical psychodiagostic instruments, such as the MMPI and the Rorschach, with measures of normal human attributes. To

be sure, these two types of assessment instruments are not mutually exclusive, as clinicians are often interested in client scores on both types of tests, and psychodiagostic tests have been used (and misused) in employee selection. Nevertheless, there are many inventories which were constructed and intended for use with normal populations, and for purposes to include employee selection, insofar as they measure job-relevant attributes. Among these are the Edward's Personal Profile, the Hogan Personality Inventory, and the various Gordon's personality and values measures.

While many of these noncognitive instruments purport to measure attributes which would logically seem to be related to job performance, they are seemingly impossible to validate on a content basis. A criterion-related study, which demonstrates the link between test scores and job performance statistically, seems the appropriate strategy.

Therefore, this paper will present the results of three criterion-related validation studies which included noncognitive tests. All studies were similar in several respects: They were to identify tests to be used in a Civil Service selection procedure which had previously included a written ability/aptitude test, but no noncognitive test; the study was for a large job class with an "open" (not promotional) candidate group; and several nonability-based dimensions of job performance had been identified during criterion development.

The first study was for Commercial Service Representative (CSR). CSRs perform activities related to the processing of billing and other financial records and provide information to customers about water and electric service. The job requires extensive telephone contact, and some public counter interaction, with a large number of customers on a daily basis, so possible use of a noncognitive test seemed appropriate.

A research test battery, which included ten abilities-based tests and the 212-item Clerical Potential Inventory (a variation on the HPI) was administered to a random simple of 93 CSR incumbents, and the incumbents were each rated by their first- and second-level supervisors.

Twelve dimensions of job performance had been identified for this job. Ratings were factor analyzed (principal components, varimax rotation) and three factors emerged: Ability (Quantity of Work, Quality of Work, Problem Analysis, Judgment); Service Orientation (Clarity or Oral Communications, Manner of Oral Communications, Sensitivity to Others, Patience, Cooperativeness, Customer Orientation); and Dependability (Supervision Required, Reliability).

Scores on each of the four scales of the Clerical Potential Inventory (Rehability, Stress Tolerance, Service Orientation, Clerical Potential) were correlated with each of the three factors. In predicting Service Orientation, three of the four correlations were statistically significant; in predicting

Dependability, none were (although two were very nearly so); and in predicting ability, as anticipated, all correlations were near zero. The prediction of Overall job performance by the combined cognitive tests was .28 ($p<.01$), by the combined noncognitive tests it was .24 ($p<.01$), and by their total was .35 ($p<.001$). Predictiveness of the Service Orientation factor by the noncognitive tests was notably high ($r=.31$; $p<.01$) and, as this was a particular concern to the employer, the Clerical Potential Inventory was retained for use in the selection process (as part of the 50% weighted written test component).

The second study yielded less favorable results. It was for the job class of Meter Readers, who travel to customer locations according to an assigned route, read meters which indicate electricity and water use, and record results. Because this work must be done without direct supervision, and considerable initiative in attaining some readings is required, it was thought that noncognitive, personal characteristics might be predictive of job performance. For this study, the 198-item Prospective Employee Potential Inventory (another HPI variant, which is scored on Reliability, Stress Tolerance, and Service Orientation) was selected for use.

This questionnaire, along with nine ability-based tests, was administered to a random sample of 94 Meter Readers, and job performance ratings were collected by two levels of supervision above each study participant.

While many of the ability-based tests correlated at a statistically significant level with the performance criteria, none of the noncognitive tests did. In fact, all of these correlations were near zero, and there was no consistent pattern of positive or negative correlation.

The third study which included research on noncognitive tests was Traffic Officer. Traffic Officers direct traffic at busy intersections, ticket illegally parked vehicles, and arrange for the impounding of repeat violators. This job requires both extensive working without supervision and contact with the public, most often in sensitive situations.

The noncognitive test used in this research was the Personnel Decisions Employment Inventory, which includes the two empirically keyed scales of Job Performance and Tenure. This questionnaire was administered to 93 incumbents, along with ten ability-based tests. In this study, validation results for even the ability tests were considerably weaker than in the two previous studies; and the noncognitive tests did not correlate with job performance at a statistically significant level, even though several of the criterion factors would logically be related to them (e.g. Reliability, Willingness to Work, Initiative). Specifically, the Job Performance scale correlated near zero with all job performance ratings, with about equal numbers of positive and negative correlations; the Tenure scale correlated positively in 13 of 14 cases, and two of the correlations were statistically significant, but this was not deemed a sufficiently strong result to warrant use of the scale for selection.

This paper has presented results of three criterion-related test validation studies that included research on noncognitive tests. Results presented have been mixed. Rather than speculate on the reasons for these mixed results, or cite the evils of sampling error, I will conclude with a plea for more much-needed research in the area of noncognitive test use in employee selection, which should ultimately provide clarification.

* * * * * * * * * *

LOOKING FORWARD: RESEARCH DESIGNS THAT LEAD TO INNOVATIVE TESTING

Donna L. Denning, Personnel Research Psychologist
and Frances Aiello, Personnel Analyst
City of Los Angeles

The City of Los Angeles employs nearly 50,000 people who fill nearly 1,300 job classes. Tailoring examinations to each of these classes can be cumbersome and time consuming. Research studies with designs that lead to innovative testing can not only enhance the job-relatedness of the testing devices, but also make possible the use of testing devices that were formerly not available.

The small size of many job classes can be seen as an impediment to large scale research projects. However, by grouping classes in terms of salient, job-related dimensions, the large scale research project becomes feasible. Three applications of this type of research design will be discussed.

In the first study, the target group is all first-level supervisors. The attempt is to identify a paper-and-pencil test of supervisory potential to uniformly examine for the supervisory component of these jobs. The study is based on the premise that there is a supervisory component common to all first-level supervisory jobs, exclusive of the technicalities of the job. Certain job activities are performed by all first-level supervisors, regardless of the type of work they supervise. Job analyses have continued to substantiate this premise. This component includes tasks such as assigning work, scheduling work, and evaluating employees.

The first step was to identify the target classes. This was done by reading all of the City's Class Specifications and assigning each class a code (0 = non-supervisory, 1 = lead worker, 2 = first-level supervision, 3 = management). The criteria used to determine first-level supervision included 1) were formally

supervising employees (e.g., assigning work, monitoring progress, evaluating performance, approving time off), and 2) had no supervisors reporting to them. All classes which were coded "2" were eligible for participation in the study. This translated into 293 job classes with 3,847 incumbents.

In order to develop a specific and detailed understanding of the nature of supervisory work, several sources were used. Various City of Los Angeles job analyses of supervisory classes, a large-scale study of supervisory classes done in 1980 in the City of Los Angeles, as well as published, generic job analyses of supervision were studied.

The next step was test identification. Two commercial tests, developed specifically to assess supervisory potential, were chosen. In addition, a test constructed by the analyst, similar to the testing done for supervision on current examinations, was chosen for use.

A performance rating form was developed specifically for use in this research project. Supervisors, one level above the sample group, were used during this phase to identify information necessary to construct this form.

The research tests were administered to 200 randomly selected, first-level supervisors during a three week period. The participants represented all major departments and all job types. Job Performance ratings were collected on each participant, two ratings per participant.

Ultimately, we hope to identify a battery of tests which demonstrates a positive, statistically significant correlation between test scores and job performance. This will show that the chosen test is valid for use as a selection device to assess the supervisory component of the job classes. Analyses will also be done to evaluate adverse impact and to assure test fairness.

This test can potentially be used as an examination section, along with other tests tailored specifically to the job, for all of the classes in the study group.

For the next two studies I will discuss, I will not go into as much procedural detail, but will concentrate on the underlying rationale for each.

The second study, on the drawing board only at this time, is a natural follow-up to the supervisory study. This study's focus is on the establishment of an Assessment Center for use in selection of candidates for upper management/executive level positions in the City of Los Angeles. The City has used Assessment Centers, specific to the job, as a selection device in the past, however, sparingly. The limited use was due to the considerable time and resource commitments necessary to develop Assessment Centers specific to particular jobs.

Use of an Assessment Center as a selection technique provides a unique opportunity for the reliable, valid evaluation of a wide range of managerial skills and abilities (e.g., leadership, organizing and planning, decision-making) which are not readily evaluated by other means. Via establishment of a generic

Assessment Center (neutral with respect to job content), a single center can be used for selection into any position requiring a comparable level of these managerial skills and abilities, thus making Assessment Center use a feasible and cost-effective means of making such selections.

This approach is preferable to intermittent use of stand-alone assessment center exercises for several reasons:

(1) It provides for an a priori, comprehensive determination of the integration of use of assessment into existing Civil Service System.

(2) It streamlines the job analysis process, and eliminates redundancy by including a single large-scale analysis of all appropriate classes.

(3) It eliminates reliance on stand-alone job simulations. Assessment Centers have consistently demonstrated validity when used in a variety of organizations; but validity data on individual exercises has been less encouraging. By developing a generic Assessment Center for use, agencies can benefit from the use of a typically valid predictor, and eliminate the recurring costs (time and resources) associated with developing individual exercises.

(4) It allows for extensive study, construction, and review of assessment exercises by a limited number of analysts, and the ultimate designation of a limited number of exercises for use, rather than requiring an inefficient procedure of various analysts constructing various, similar exercises for use with different classes at different times.

(5) Use of neutral content permits the assessment of "pure" managerial skills and abilities, unconfounded with job knowledge, technical ability and/or specific previous job experience. (These may be critical attributes for a given position; but they are, at best, inefficiently measured and, at worst, inaccurately measured in an Assessment Center. In this arena validation studies are lacking.)

(6) Only in a complete Assessment Center is the full strength of the method permitted to operate: each candidate is seen performing by multiple trained assessors in multiple situations which tap multiple job-related attributes. Written tests and/or an interview may also be used. Final evaluation is based on the integration of all these information sources.

In the last study to be discussed, focus is shifted from upper level classes to entry level classes. The basis for this study stemmed from the realization that many entry level classes such as Tree Surgeon Assistant, Airport Information Aide and Parking Attendant, call simply for a basic skills assessment - reading, writing, and arithmetic.

Currently, when an examination is needed for each class, the analyst learns about the job, identifies areas which need to be tested, and either writes items or uses previously written items (reviewed by Subject Matter Experts). For many of these classes, specifically many entry-level classes, analysts find over and over that the areas needing to be tested are the same - the basic skills mentioned above.

From a strictly content-related validation standpoint, job analyses for these jobs identify a common denominator of basic skills necessary for successful performance of the job tasks. By capitalizing on this, a single test battery can be developed for use in the examining process for all classes in the target group to assess these basic skills.

Though the study can be done using a content validation approach, the mere numbers involved make a statistical study feasible also. This is especially desirable given the generic nature of the testing.

In this paper, three research studies were discussed: a Supervisory Study, a General Management Assessment Center Study, and an Entry Level Basic Skills study. The Supervisory Study is near completion, and the other two studies are in the developmental stages. By conducting research studies using designs such as those discussed in this paper, agencies can benefit from improved measurement and increased efficiency in their examination processes.

* * * * * * * * * *

SAN DIEGO COUNTY CAREER DEVELOPMENT ASSESSMENT PROGRAM - AN AFFIRMATIVE ACTION PROGRAM TO IDENTIFY AND DEVELOP EMPLOYEES WHO HAVE DEMONSTRATED MANAGEMENT STRENGTH

Del Boenrer, Senior Personnel Analyst
County of San Diego

The San Diego County Management Academy is an employee development program which utilizes the assessment center process to identify employees at all levels within the permanent County workforce who have demonstrated superior management skills. Having identified these employees, the program provides for developmental exercises and training to make them extremely competitive in promotional examinations for supervisory and management positions. This paper addresses the general program concept and implementation. Information relative to the validation of the assessment center exercises will be provided by Mr. Richard Joines, Management Personnel Systems, Inc., who served as a consultant to the County in the development of the selection process.

Since 1977, the County of San Diego has operated under a consent decree with Department of Justice oversight insofar as personnel training and selection is concerned. In April, 1985, the County implemented an Affirmative Action Plan

which included a requirement to provide an employee development program for the purpose of enhancing minority selection opportunity for management positions. This plan recognized that the County had made significant progress in the employment of minority persons but that this progress was primarily at the lower levels within the County hierarchy. At a September conference with the Board of Supervisors, the Director, Office of Employee Services, proposed an employee development program which embodied many of the concepts present in the current program. In addressing the Board, the Director advised that any worthwhile program would require a significant investment in time to research and implement and she asked the Board to make a five-year commitment to the program. The Board approved the program with its five-year concept and in January 1986 we formed a small staff to carry the work forward.

One of the first tasks for the Accelerated Career Training Staff (ACT), as we call ourselves, was to develop an implementation plan. This plan was charted out using a modified PERT (Program Evaluation and Review Technique) diagram and incorporated the timetables established by the Board. This implementation plan proved invaluable in keeping our thinking clear and the program development on track.

Following the development of the implementation plan, we undertook a widespread search for information concerning management selection and development programs. This search included review of technical periodicals and books, computer data banks, and survey of California counties and large cities, as well as major employers in the San Diego area. This search was organized and documented into a computerized reference file wherein the listing containing authors with titles of their works was 34 pages long. This reference file was also arranged by topic and served as our primary defense against informal challenges which were to surface from time to time.

Concurrent with the research phase, we tested the original model which had been proposed to the Board of Supervisors. This model contained provisions for "deep classes" and guaranteed promotion during participation in the program. We discussed this model with a variety of department heads, ethnic organizations, representatives of various classes, and finally with boards and commissions which had been established by the Board to advise regarding affirmative action matters. We soon learned that deep classes, i.e. classes which spanned several pay levels, and automatic promotion were an anathema to most appointing authorities and we would have very potent opposition if the program were to be implemented with those features. This caused us to take a hard look at what we were trying to do and we found that our team had differing concepts of what the goal of the program really was. We went back to the drawing board and after some brainstorming we agreed that the goal of the Management Academy should be:

> To develop a pool of exceptionally well qualified in-house management candidates to facilitate the meeting of affirmative action plan hiring goals to management classes.

With this goal in mind, we now needed to go back to the program sponsor, the Director, Office of Employee Services, and convince her that the basic concept of the overall program must change if we were to have a successful program. Our presentation to the Director took the form of a force field analysis and at its conclusion the Director reluctantly agreed to a program which did not include deep classes or automatic promotion and we were free to redesign the program along its present lines.

As presently configured, the Management Academy Program is a developmental program for permanent County employees which makes no provision for promotion. It is structured into two competitive groups. The middle management competitive group is for journey level professional, technical, public safety, and administrative classes up to, but not including, the deputy director level. The entry level competitive group is for employees in trainee and entry-level professional, technical, and public safety classes, and para-professional, clerical, crafts, construction, and maintenance classes. There is no effort to screen out applicants other than to promise applicants that the successful individuals will be required to complete a very demanding program while doing their regular work. There are no minimum levels of education or experience required of applicants. We give the program wide publicity during recruitment periods and actively seek minority participants. The application process is deliberately designed not to require supervisor or department head recommendation, a requirement that ethnic organizations felt would work to the disadvantage of their members.

Our research confirmed that the process we wanted to use to select candidates for the program should be the assessment center. We needed to ensure that any process used was completely valid and could withstand any challenges so we designed a request for proposal (RFP) and consultant selection criteria based on our needs and mailed the RFPs to a list of nationally prominent authorities. Our search ultimately led us to Mr. Richard Joines of Management Personnel Systems, Inc.

Mr. Joines, in cooperation with the ACT staff, designed and conducted the task analysis of the target class level, sample size 229, followed through with the dimension and exercise identification, designed the pre-screening in-

basket and assessment center exercises,. trained assessors and assessment center administrators, and supervised the administration of the first two assessment centers. He will provide validation information relative to the pre-screening in-basket and the assessment center exercises.

We chose to test our concept on the middle management competitive group, recruiting for applicants during the month of April, 1987. We received 475 applications for the program. When we held the pre-screening in-basket in May, 368 applicants appeared. These in-baskets were hand scored by the staff using a 4 or 5 level narrative evaluation, and a minimum pass point was set. To ensure adequate ethnic minority representation in the Academy members, we selected by ethnic group, except for caucasian, at the rate of 1.5 times their representation in the County employee workforce. (County employee workforce representation equals or exceeds County-wide workforce representation.) This selection rationale worked quite well except for the native American group which had too few participants and these failed to meet the minimum pass point. From the 368 in-basket participants we selected 60 candidates to go on to the assessment center exercises.

In July and August we conducted five one-day assessment centers of 12 candidates each. These began with two six-person leaderless group discussions, followed by individual exercises which included both oral and written reports, and subordinate counseling exercises. Out of the 60 assessment center participants, we selected 36 management candidates who were enrolled in the Management Academy. They represent a wide variety of classes such as senior physician, senior investigative specialist, social services administrator, and analyst II. Both successful and unsuccessful candidates have access to a companion Career Counseling Workshop to assist them in career appraisal and planning.

The assessors for these assesment centers were senior department managers trained by the consultant. Not only did the assessors acquire new skills, they also found it refreshing to have the opportunity to preview some of the County's brightest employees. In addition, they have begun to use assesement center techniques in making their departmental personnel selections, a side benefit for the County in its quest to improve the quality of the workforce. Feedback to the ACT Staff indicates that employees have reacted positively to this revised selection process as opposed to the regular departmental interview.

As mentioned earlier, there are three formal boards and commissions appointed by the Board of Supervisors which are charged with affirmative action responsibilities. These boards were interested in the selection process for the

Management Academy and had expressed concern that the process might be discriminatory to their particular constituencies. As a result, they requested and were granted the opportunity to observe the in-basket and other assessment center exercises. One such observer even conducted exit interviews of participants. As a result of their observations, the boards and commissions are satisfied that the process is eminently fair and have given us their full support.

Like the assessment center, the Management Academy is not a building or location. Rather, it is a concept based on adult learning theory as described by Malcolm S. Knowles in his works on the subject. In his works, Knowles indicates, among other things, that adults need to be self-directing, that learning should be centered on real-life situations, that group members are a rich resource in learning, and that group atmosphere should be cooperative, informal, democratic and active. These theories and principles are in the forefront of the Academy design. Basically, the Academy program consists of four distinct elements: first, the demonstration of communication competencies; second, study in the theory of supervision and/or management; third, study in County-specific coursework; and finally, job-related learning experiences in County-unique activities.

These requirements were identified through the task analysis, assessment center results, feedback from County executives, personnel department evaluations of management recruitments, and individual participant questionnaires which had been completed during the assessment center process. Each candidate's program is laid out in an Individual Development Plan (IDP) negotiated between the candidate and senior department management, in many cases the department head. The ACT staff is a resource in this negotiation and periodically follows up with the candidate to monitor progress and offer assistance.

Completion of the individual requirements of the plan are in addition to the candidate's own workload and may take up to 2 years to complete, although some candidates appear to be on the way to completing their particular plans in less than 12 months.

The plans are comprehensive and require dedication to complete. For example, the first element of the plan, demonstration of communication competencies, requires three specific instances each of demonstration of oral and written communication, and oral presentation skills and sets specific conditions for successful demonstrations. In addition, it provides criteria against which these demonstrations are to be evaluated. County executives now have a clear responsibility in ensuring that future County managers are effective communicators.

The second element of the plan, theory of supervision and management, relates to knowledge which can be acquired through study in supervision or management courses at a college or university, or in County-specific courses on the subjects. The intent is to ensure that future managers have an exposure to the theories and techniques essential to be effective in these areas. Some candidates are fulfilling a portion of this requirement by completing a Certificate in Management Program at a local university.

There are many County-specific knowledge areas with which all managers should be familiar. These are covered in the third element of the program, County-specific coursework. This covers such topics as budgeting, personnel issues, and discipline. Some candidates are highly skilled in these areas and it was prudent to use those skills to the benefit of the remainder of the candidates. So, after providing training-for-trainers to the skilled candidates, we set up a series of classes on County related coursework which is proceeding on schedule using those candidates as instructors.

It was impossible to set up coursework for all types of desirable County-related experience so the final portion of the IDP requires completion of job-related learning experiences. These experiences are intended to expose the management candidates to [illegible] activities most managers are likely to face in their daily routine. They include such activities as "serve as or assist the management representative in an employee appeal to the Civil Service Commission" and "serve as or assist appointing authority's representative in a performance appraisal appeal." Department executives have been exceptionally enthusiastic in supporting this aspect of the IDP and have created a wide variety of their own job-related learning experiences which are enhancing the development of the management candidates.

The training being provided to the management candidates comes from several sources. We're using the local colleges and universities, the Regional Training Center (supported by a consortium of local governments), County trainers, contract trainers, and management candidates themselves. This diversity seems to be filling the needs of the individuals without placing the candidates in a rigid training schedule or format.

An important side benefit of the Management Academy is the networking it has provided to the candidates. They have established their own networking organization with monthly luncheon meetings. These meetings usually involve interesting speakers, further adding to the management candidates' knowledge about County operations. In addition, the candidates have been invited to join other County organizations, an opportunity which might not have been available but for the Management Academy. Finally, most have gained increased access to senior management and the executives in their respective departments. With this exposure, some candidates have noticeably blossomed and see a real potential for the realization of their ultimate goals.

We asked ourselves early in the planning process how we would determine success of the Management Academy. Our answer was to establish a control group of employees approximating as nearly as possible the management candidates in class, age, service with the County, ethnicity and gender, and to periodically compare the promotions within both groups of employees. Although we selected our first management candidates less than 12 months ago, 24% have been promoted at least once, some twice, while the promotion rate of the control group is 5.8%. Needless to say, the effect on the group as a whole has been electrifying. When the management candidates meet as a group one feels the enthusiasm and excitement present. They've convinced themselves that there is no challenge that they can't conquer, if not individually, then, together. We have been truly successful in fulfilling our goal -- we have identified for County executives the best and brightest employees for future management appointments.

This success has not gone unnoticed. County executives now seek out challenging assignments for management candidates, and other employees are competing for acceptance into the Academy. They see it as a way to gain recognition of their capabilities and enhance their opportunities within County government.

When we began this program in January, 1986 success was not a foregone conclusion. There have been many places where we might have taken a wrong turn, or alienated an important supporter. So, why has this program succeeded? First and foremost, I believe that the Director's sense of timing coupled with the persuasiveness of her arguments to the Board of Supervisors were the key factors in getting the program off to a flying start. The Board was ready for an initiative having just approved the Affirmative Action Plan early in the year, and the Director was ready with a plan which gave focus to their desire to demonstrate that they were willing to pay a reasonable price for results.

Next, the design was right for the County of San Diego. Our research paid off by giving us the benefit of the experience of others and applying that experience to the circumstances in the County. Although we found no program like the one we've implemented in San Diego County, we did find developmental assessment centers that we liked and management development programs that we borrowed from in the design of our own program. Not only was the design right, our program planning process kept our efforts focused. We used a modified PERT (Program Evaluation and Review Technique) which gave us a clear visual picture of where we were and where we had to go. It was easily changed, yet once we set out our goals and process, we found that few changes were necessary.

In addition, we developed a broad base of employee and management support by including employees and managers in as many ways as we could. We used a task analysis survey which involved a large number of middle managers, involved managers in identifying dimensions and as assessors, gave management and executive briefings, briefed employee associations and advisory groups, prepared and distributed literature about the program, and aggressively recruited for participants in the program. This broad base of participation coupled with our ability to respond immediately and effectively to challenges to any part of the program helped us to forestall any formal challenges.

Finally, and certainly not the least consideration in the success of the program, has been the quality of the candidates that we identified and are developing. The management candidates themselves are the best evidence of success that we have. Their willingness to undertake the challenges of the program is inspirational. We will see many of them in most responsible positions in the County organization within the next few years. Certainly, we have fulfilled the goal of the program.

* * * * * * * * * * *

WE DID IT BEFORE - WILL WE DO IT AGAIN?

(Will selection specialists react constructively if we have a financial depression?)

Ted Darany, Employment Division Chief
San Bernardino County, California

There appears to be a dramatic challenge before us: a significant downturn in the nation's economy. This paper addresses 1) the potential of this actually happening, 2) how should selections specialist respond generally, 3) specific suggestions for selection and general personnel practitioners.

A DOWNTURN IN THE ECONOMY

There's been much discussion and several national best selling books on the subject of an economic recession or even depression. Crashes, recessions, panics and depressions have been predicted over the past 20 or 30 years. Why should we worry now?

It would seem that there has been one significant change in our nation's economic health (and that of most of the rest of the world as well) in the last two decades: DEBT. Debt is the "wild card" in our current deal of recession/depression. We've all heard that debt has reached levels never before seen in this country. What seems to be critical about this build-up of debt is the way it may interact with a recession. Currently, we are in a generally strong economic period. However, the economy has a long standing pattern of ebbing and flowing with the general business cycle. The possibility that we have become wise enough to totally avoid a recession seems most unlikely. In retrospect, a recession may be seen to have been caused by any number of factors such as high interest rates, increasingly scarce commodities or employee talent, or a catastrophic event -- natural or man-made. But most recessions seem fairly easily explained as a natural course of ending the up-move in a particular business cycle. Currently we are in what may viewed as the longest peace-time non-recessionary period of this century. It does not seem unreasonable that this positive economic period may end reasonably soon with the onset of a "normal" recession. But with the onset of a normal recession this time, we have to consider the debt wild card.

During a recession, revenues by government typically drop while demands on governmental services increase. This always puts a squeeze on available money for businesses to operate, often pushing our economy further into recession. This time, with debt so high for government, businesses and individuals, this spiraling down of the economy has a chance to accelerate out of control. It's not my view that recessions and depressions just mystically happen. It may appear that they're caused by a combination of bad luck (several unlikely events occurring at the same time) as well as bad decisions. Of course, the bad luck and bad decisions are only clearly bad in the wisdom of retrospect. Given the pressures our debt will place on us during a recession, it may be expecting too much good luck and too much perfection in the decisions of our policy makers for this normal recession not to accelerate into a full-blown depression.

We need to ask ourselves how severe might this new recession/depression become? Let's focus on unemployment statistics since they're generally the most important to personnel practitioners. Currently, the unemployment rate is approximately 5 1/2% nationally. During a normal recession, the rate might rise to 8% to 11% (1982-83 averaged 9.5%). That difference might not seem large but it reflects an enormous problem for our nation. It drastically reduces revenue for all levels of government and corresponds to real trauma for a large proportion of our population. But this is only a recession. If our massive debt results in an extraordinary business contraction, unemployment might exceed 20% -- levels not seen since the 1930's in this country. Such an outcome would have significant impact on virtually all of our institutions. If we have such a depression, there will undoubtedly be similarities to the one we had in the 30's, but there will also be major differences. It seems fairly certain though that if this period extends more than a year, it will have a self perpetuating influence on our business, government, and social attitudes. At a certain point, many of us will resign ourselves to the situation and quit fighting it. And that may be as big a problem as the debt which triggered it, since some confidence in the future seems essential to actually beginning to move out of a depression.

In the rest of this paper, I will focus on how personnel practitioners can be a positive force in lessening the impact of any upcoming recession or depression by preparing ourselves for it and working against it if it arrives.

HOW SHOULD SELECTION SPECIALISTS RESPOND? - STRATEGIC PLANNING

One method which may be useful to many of us in preparing for significant change is strategic planning. In this context, the most important issues are evaluating the extent of the possible downturn, developing planned responses for each type of downturn, and assessing current resources vs those likely to be available and needed during each type of downturn.

How do we assess the extent of an economic downturn? I will suggest two statistics: unemployment rate and help wanted advertising "lines". Of the two, the unemployment rate is much more widely known and available but it tends to be, at best, a "trailing" statistic. The help wanted statistic has been a better indicator of trend changes in the economy. That is, it's more sensitive to an economy which has peaked and is starting to turn down. For example, if we see help wanted advertising lines start to decrease while unemployment stays at a relatively low level, it may be that the economy has already started to slow but that it has not yet shown up in the unemployment statistics. It is similarly useful as an indicator that the worst may be over in times of a recession. As we are moving from a recession towards a depression, unemployment will pass through 10% to 12% and move into the teens. Help wanted advertising will drop precipitously. No bell will go off signifying "the depression" has started. But that sort of trend would certainly suggest the possibility. While these two national statistics are the most reliable and best indicators of our national state of health, they may conceal what's more important to us: the state of health of our local economy. Let's look at some possibilities.

Many regions of our country have had extraordinarily severe economic periods over the last 15 years. It is not unfair to characterize the situations in some of these regions as a depression. The regional problems have become so popularized that they have been given nicknames, such as "rust belt" recession, "energy patch" depression, and of course our farmers have been through two major down cycles during this time. Family businesses were lost, unemployment skyrocketed, local and state governmental agencies were severely pressed. This all occurred during periods when the national economy was relatively free from recession and absolutely not in a depression. The effectiveness of our response to economic downturn depends a great deal on an understanding of the scope of the problem. Our response needs to be tailored to the scope of the problem. But there's one more factor to consider as well: the health of private vs public sector activities. There are situations where a region may have a dramatic change in the economic health of either the private or the public sector while the other sector remains relatively stable. For example, there have been periods in which the aerospace industry had dramatic upturns and downturns while the governments in the communities significantly affected by aerospace employment remained relatively stable. On the other hand, there have been times when government agencies have had severe revenue reversals while the private sector remained on a stable course. Again, the point of this analysis is to tailor our strategy to the circumstances so that it may be most effective. One example here may be helpful. After the "taxpayer revolt" in 1978 in California, the State's local governments had a dramatic downturn in their revenues. However, the State's general economy remained satisfactory. Therefore, it was possible and effective to wage an aggressive out-placement effort of current employees who might otherwise be laid off. This would save taxpayer dollars in unemployment insurance, reduce the trauma of the to-be-laid-off employee, and move a potentially productive person to a useful job in the private sector. Such a program had lasting benefits for the entire community when compared to the simpler but more wasteful methods of staff layoffs. Obviously, these methods would not have been as useful in the energy states during their recent regional depression since both public and private sectors have suffered tremendously. And in this light, it should be admitted that any severe trend in one of the two sectors will eventually impact the other sector. But it seems useful to reflect that even without a national depression, local problems come along to all of us sooner or later. So it's useful to consider what sorts of resources we will need and compare that with what is likely to be available during such a downturn.

SPECIFIC STRATEGIC SUGGESTIONS FOR RESOURCE DEVELOPMENT

This section focuses on the development and use of resources likely to be especially beneficial during a period of economic recession or depression. However, many of these ideas have high usefulness even during more normal times. The suggestions may be broken into three categories: developing cost containment resources, developing external resources, and developing skills effective in dealing with bad times.

Focusing on cost containment resources first seems pretty sensible. We're all trying to hold the cost of government to the lowest possible level. However, in the context of this paper, what I'm recommending is: don't wait for it, that is, the depression to occur. The primary concern to focus on is prioritization. What is it about our organization that our clients most need? For most of us, our clients are other governmental departments in our system or perhaps the

general tax-payer. What we should really set out for ourselves, are those absolute essentials without which we wouldn't be doing our job. Doing this first, will allow all of the other resource development to be much more effectively adapted to our needs. A second cost containment consideration is simplification. This simplification should follow directly from our prioritization. That is, we should develop a simplification plan for what we would do if required to make our organization more streamlined or smaller. This may entail getting rid of some favored or pet programs that might have been a high priority to a high official, but which during tough economic times would almost be an embarrassing frill. A third point under cost containment is automation. It seems we're all rushing headlong towards automation right now. So this suggestion is really: automate effectively. While that seems an obvious suggestion, it seems that many organizations have not automated that effectively. Some of us have automated activities in our work for which automation wasn't a particular benefit. Or in some circumstances, some of us have automated with approaches that were ill-suited to our needs -- sometimes resulting in even more staff time necessary to accomplish a task than before we automated. Suffice it to say that in automation, as in most endeavors in life, there is a very wide range of solutions, from solutions which are so inefficient for a particular task as to make the result less satisfactory than before the automation, to solutions which are so well suited as to save significant amounts of staff time, money, or to produce significantly better services to the public. Obviously, we should seek the latter. There is a wide range of resources available to assist us in deciding which approaches work best to satisfy our automation needs. Many of us have not always reviewed these information resources before we made our automation decisions. It's never too late. It may very well be that the most practical solution would be to throw away a previous automation solution and replace it with one which is truly effective. The essential here is to acquire the specific knowledge necessary to know our needs and know the available solutions to those automation needs to make as ideal a fit between need and solution as is possible. The fourth suggestion provided here would be to develop a plan for how our organization could grow smaller. That is, actually plan to grow smaller. How would we produce the work required of us based upon the priorities that we established with a steadily decreasing staff size. That exercise, itself, will probably do wonders to refocus us on our priorities.

The second major resource area I would suggest is to develop external resources. First, I would recommend active pursuit of "helper" networks. By this, I'm suggesting such organizations as IPMAAC, PTC, and IPMA. These are organizations which may be valuable for information, training, or finding others with similar problems. Second, I would suggest learning more about becoming a member of specific-purpose organizations. One type of specific purpose organization is the regional consortium such as WRIPAC, GLAC, and MAPAC which are dedicated to developing cooperative solutions to problems in the selection field. These are active on-going organizations which meet periodically at rotating locations in their regions. They have been particularly beneficial in the development of cooperative training and also been productive in several specific joint projects. Members represent public agencies in their regions, but visitors are generally welcomed at their periodic meetings. A second type of organization is one which has been formed to cooperatively meet a specific need. Examples are the

Cooperative Employee Search Association (CESA) and the Western Region Item Bank (WRIB). WRIB, the first of these organizations, was formed in 1981 with 18 members with a goal of sharing test question resources across its membership. It has been useful enough to the selection field to have grown now to 103 agencies in 19 states. Both of these specific organizations are administered by the Employment Division of San Bernardino County, California.

The last suggestion offered is skills development, focusing on skills which may be effective in dealing with bad times. While specialists in selection may already have some of the skills, personnel practitioners in other areas can readily acquire them, too. Moreover, very few selection specialists have developed these skills extensively. The suggested skill development areas are: job stress counseling, out-placement counseling, job search, and career development.

If we have bad economic times in our future, preparation now should help us to get through those times. If we're fortunate enough not to have to endure severe economic times, the suggestions provided in the latter section of this paper probably will lead us to a more effective and purposeful organization.

Suggested Readings:

1. A Strategy for Resource Allocation in Public Personnel Selection, Charles F. Sproule. Presidential address presented at the June 1979 Annual Conference of the International Personnel Management Association Assessment Council, in San Diego, CA.
2. Extraordinary Popular Delusions and the Madness of Crowds, Charles Mackay, LL.D. Farrar, Straus and Giroux, New York.
3. Strategic Planning in an Information Economy, Michael Rogers Rubin. Information Management Review, vin2, Fall 1985.
4. Shaping Strategy: Tie Personnel Functions to Company Goals, Gerald R. Ferris; Dan Curtin. Management World, v14n1, Jan 1985.
5. Down-Sizing Your Company to Meet New Realities, B. Charles Ames. Industry Week, v224n4, Feb 18, 1985.
6. The Search for Quality in the Face of Retrenchment: Planning for Program Consolidation Within Resource Capacities, Thomas R. Mason. Paper presented at the Annual International Conference of the Society for College and University Planning (19th, Cambridge, MA, July 10, 1984).
7. A Strategic Plan for the Oregon State System of Higher Education, 1987-1993, Oregon State System of Higher Education, Eugene, July 18, 1986.
8. Saving Millions Through Judicious Selection of Employees, Charles B. Schultz. International Personnel Management Association, Volume 13, No. 4, Winter 1984.
9. Computer Applications to Personnel (Releasing the Genie -- Harnessing the Dragon), Theodore S. Darany. International Personnel Management Association, Volume 13, No. 4, Winter 1984.
10. A Bridge Collapse and Personnel Selection, Jame P. Springer. International Personnel Management Association, Volume 13, No. 4, Winter 1984.
11. Staffing the Public Service, Albert P. Maslow, Ph.D., Book Crafters, Inc., Chelsea, Michigan, 1983.
12. Determinants of Work Force Reduction Strategies in Declining Organizations, Leonard Greenhalgh; Anne T. Lawrence; Robert I. Sutton. Academy of Management Review, 1988, Vol. 13, No. 2, 241-254.

Criterion Related Validation Using Two-Way Validity Generalization

Walter G. Mann, Jr.
U.S. Office of Personnel Mangement

Washington, D.C.

Validity generalization (VG) is usually used to analyze and summarize the results of other people's validity studies. In contrast, this report describes the use of VG procedures in an in-house validation of a test battery. VG was used in a criterion-related validity study for the purpose of estimating situational specificity and obtaining assurance that the test battery does not have differential validity over (a) 39 jobs or (b) 9 job sites. VG did this without necessitating a wait of 10 or 20 years to collect a large enough N for each job title and job site. Use of the two-way VG analysis provided evidence concerning appropriate differential use of the test battery by job and job site.

METHOD

Nine tests, which are described in Table 1, were validated at nine naval installations that train and employ apprentices for federal, blue-collar, trade and craft positions. All jobs in the study require an initial four-year apprenticeship that leads to jobs such as welder, painter, mechanic, and boilermaker. Mean course grade was chosen as the criterion, largely because the key stumbling block in apprentice training is performance in classroom courses.

Validity coefficients for each of the nine installations were analyzed by the Schmidt-Hunter (1980) interactive validity generalization procedure. Validity coefficients for each of 12 jobs were analyzed with the same procedure. The 39 job titles were reduced to 12 because only 11 had N's as large as 25. The other 28 jobs were grouped and became the "twelfth job" in the analysis.

Validity coefficients were corrected for restriction in range and criterion unreliability, but not for test unreliability. The standard deviation of validity coefficients was corrected for sampling error, for predictor and criterion unreliability, and for restriction in range. The actual distribution was used for correction for restriction in range; otherwise, assumed distributions were used.

RESULTS AND DISCUSSION

For the VG analysis across job sites (Table 2), situational specificity (100% minus the percent variance explained by all artifacts) was low except for Test 102D (simple arithmetic computation). For the VG analysis across jobs (Table 3), situational specificity (SS) was generally low. Surprisingly SS was highest (42%) for the Weighted Total Test. Even this amount of SS can be tolerated because the SD of the estimated true validity was only .128 and the bottom tenth percentile was a very acceptable .695.

[1] A complete validity report is forthcoming. I will be happy to furnish a copy to any interested person.

For both VG analyses, estimated true validity was lowest for Test 100A and highest for Test 102C. In general the validity results for the two VG analyses were quite comparable.

Regression analysis or analysis of variance could have been substituted for VG in the present study. In fact, I did a regression analysis of residuals (test score minus predicted criterion score), and the results supported the VG approach.

Tables 2 and 3 contain the results using the Schmidt-Hunter assumed distribution for criterion reliability (mean = .80). After Paese and Switzer (1988) questionned the use of the Schmidt-Hunter assumed distributions for criterion reliability, I reanalyzed my data using reliability coefficients computed for each situation. Results using actual distributions were comparable to results using assumed distributions: the estimated true validity coefficients changed at the third decimal place, while the percent of variance accounted for changed at the second decimal place. This would indicate that a naive acceptance of the Paese and Switzer results or recommendations would be imprudent.

CONCLUSIONS

The test battery is highly valid overall, at the various job sites, and across jobs. The residual situational specificity left after corrections were made is most appropriately ignored. The VG procedure appears to be appropriate for situations where one has multiple jobs or multiple job sites, or both.

References

Paese, P.W. & Switzer, F.S. (1988). Validity generalization and hypothetical reliability distributions: a test of the Schmidt-Hunter procedure. Journal of Applied Psychology, 73, 267-274.

Schmidt, F.L., Gast-Rosenberg, I, & Hunter, J.E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.

Table 1

Tests in Apprentice Examination

Test	Ability Measured	Item Type
100A	Eye-Hand Coordination	Rapidly and accurately move the hands or fingers, under the coordination of the eyes
100B	Measuring Ability	Alignment dexterity using a gauge
100CH	Form Perception	Visualize a 2-dimensional form having only seen the parts that make it up (C) Inspect drawings to see slight differences in shape, size, or shading (H)
100D	Complex Arithmetic	Compute or work with fractions and decimals
100E	Memory	Follow oral directions
102A	Reading Comprehension	Read and understand sentences and paragraphs
102B	Numerical Reasoning	Arithmetic and algebraic word problems
102C	Table Reading	Follow written directions in a table
102D	Simple Arithmetic Computation	Quickly and accurately do arithmetic on whole numbers

Table 2

Validity Summary for Nine Job Sites
(Apprentice Tests Predicting Mean Apprentice Course Grade)
(N=798)

Test	Descriptive Statistics: Mean $\underline{r}$[1]	Descriptive Statistics: $\underline{SDr}$	Percent Variance Explained By: Sampling Error	Percent Variance Explained By: All Artifacts	Parameter Estimates: $\bar{\rho}$	Parameter Estimates: SD_ρ	Parameter Estimates: Bottom 10th %-ile
100A	.161	.115	78	83	.187	.056	.115
100B	.285	.083	100	100	.510	.000	.510
100CH	.286	.091	100	100	.417	.000	.417
100D	.507	.113	46	76	.707	.076	.610
100E	.282	.131	51	82	.531	.105	.396
102A	.510	.074	100	100	.767	.000	.767
102B	.557	.108	46	94	.739	.036	.693
102C	.432	.120	50	74	.789	.112	.645
102D	.307	.192	22	29	.517	.271	.169
Wtd. Total 100	.447	.110	55	100	.666	.000	.666
Wtd. Total 102	.601	.089	56	100	.892	.000	.892
Weighted Total	.595	.097	48	100	.881	.000	.881

Table 3

Validity Summary for Twelve Trades
(Apprentice Tests Predicting Apprentice Course Grades)
(N = 798)

Test	Descriptive Statistics: Mean $\underline{r}$[1]	Descriptive Statistics: $\underline{SDr}$	Percent Variance Explained By: Sampling Error	Percent Variance Explained By: All Artifacts	Parameter Estimates: $\bar{\rho}$	Parameter Estimates: SD_ρ	Parameter Estimates: Bottom 10th %-ile
100A	.179	.089	100	100	.203	.000	.203
100B	.295	.131	71	79	.490	.099	.364
100CH	.299	.121	82	90	.417	.053	.349
100D	.493	.129	56	74	.685	.092	.567
100E	.262	.114	99	100	.463	.000	.463
102A	.423	.144	52	67	.547	.107	.410
102B	.424	.142	51	62	.536	.111	.394
102C	.410	.123	71	94	.751	.056	.679
102D	.325	.123	81	93	.529	.054	.461
Wtd. Total 100	.461	.109	82	100	.677	.000	.677
Wtd. Total 102	.575	.132	43	64	.865	.120	.771
Weighted Total	.581	.134	41	58	.858	.128	.695

[1]All validity coefficients were significant at the .01 level or beyond.

APPLICATION OF ANGOFF IN PASSING POINT SETTING FOR A SITUATIONAL INTERVIEW

Lee Wieder and Thung-Rung Lin

Los Angeles Unified School District

Abstract

This paper discusses an application of the Angoff judgmental method of setting a pass point on a very structured interview, specifically a "Situational Interview". Further modifications of the Angoff method are also discussed. These theorized modifications present a stronger rationale in estimating a preset pass point as applied to structured interviews in general.

In the personnel field, passing points are used in many different ways to help managers make decisions regarding training, promotion and selection. In the area of employment testing, personnel administrators and selection specialists are often required to set the passing point for newly developed tests. Typically, these passing points serve two important functions: 1) to maintain the minimum standards of job competencies, and 2) to select the best qualified (McClung, 1974).

There are a variety of methods available for estimating passing points for multiple choice tests. Among them, the three most commonly used judgmental methods are the Angoff (1971), Ebel (1972), and Nedelsky (1954) methods. All of these methods require that judges estimate the performance of the "Borderline Testaker", or " Minimally Acceptable Candidate (MAC)" on a multiple choice written test. However, there are no equivalent judgmental methods, as far as the authors know, documented in the literature to guide the setting of passing points for interviews. The primary reason is because of the conventional interview format, even though it may be structured in nature, it is very different from that of multiple choice written test. However, the authors believe that if the design of the interview format is so structured that it can be viewed as an "orally administered written test", then some of the above judgmental methods may be transferable from the multiple choice written test to the structured interview with minimal adaptation.

The purpose of this paper is to document an attempt in pass point setting for a highly structured interview namely, the Situational Interview (SI) (Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980), by using the Angoff method. Thus, the focus of this study is the empirical application of one of the judgmental pass point methodologies rather than a discussion of pass points in general. Readers who are interested in the broader discussion of the legal, psychometric, and professional issues relating to passing scores in employment settings should read the excellent review by Cascio, Alexander, & Barrett (1988). Readers who are interested in the available methods in pass point setting should also read the extensive review by Beck (1986). Readers who are interested in the conceptual discussions on what are the "standards and criteria" in pass point setting should not miss Glass (1978).

What is a Situational Interview ?

The situational interview (Latham, et al,1980) is an interview based on a systematic job analysis known as the critical incidents technique (CIT) (Flanagen, 1954). The incidents are collected and structured into interview questions in which applicants are asked to indicate how they would behave in given situations. Each answer is rated on a five point Likert-type scale. To facilitate objective scoring, job experts develop behavior statements that are used as benchmarks or illustrations of 1, 3, or 5 point answers, (5 being the optimum response).

The Setting

Using the critical incidents job analysis technique, thirty one situational questions and their corresponding benchmarks were developed for the classification of School Custodian for a large west coast urban school district (Lin, 1988).From these thirty one questions, two parallel situational interview forms (Forms A & B) were constructed, consisting of 20 questions each, with some overlap of questions between the two forms.

A preset pass point for the situational interview was needed for the field employment office administration of the 1987 school custodian examination. The format of the situational interview is similar to that of multiple choice tests because: (1) the situational interview format is highly structured, (2) there are precise and quantifiable benchmark answers for each interview question, and (3) the same set of questions is used in the situational interview for each candidate.

The application of one of the judgmental methods commonly used to set pass points for multiple choice tests was used for the situational interview. The Angoff judgmental pass point method was chosen due to its wide usage by personnel practitioners and the relative ease of instructing the subject matter experts (judges) on its application.

Pass Point Setting Procedures- Stage 1

The basic outline of the five steps commonly used in the Angoff judgmental method are (Livingston, and Zieky, 1982):

1) Selection of qualified judges,
2) Define "borderline" knowledge, abilities and skills,
3) Train the judges in the use of this method,
4) Collect judgments, and
5) Combine the judgments to choose a passing score.

Following the above outline, a meeting was conducted with seven highly qualified subject matter experts (SMEs). In the meeting, the Angoff method was introduced and the SMEs were instructed in its use.

After the completion of the judgments, an estimated minimum pass point was derived by averaging the SME ratings of each item and then averaging the item averages.

Results

Combining the SME's item ratings resulted in an estimated pass point of 59.28% for form A, and 61.61% for form B. As a result, the final pass point for both situational interview forms was set at 60%.

The coefficient alpha internal consistency reliability estimate for the seven SMEs was derived for both forms. The estimate for both forms resulted in an identical reliability of .73.

Discussion

How effective and useful was this preset pass point derived from the Angoff method, for the School Custodian examination? Using 60% as the preset pass point, more than 90% of the candidates passed the SI and were placed on an eligibility list. However, if the SI had not been the final test part in a multiple hurdle examination, this preset pass point would not have significantly cut down the number of candidates. Nevertheless, in this case, the SI was the final test part and the selection ratio was low. Thus, this preset passing point seems to have had little utility other than formality (Cascio,et al, 1988, p.4).

However, the real questions are: Was the pass point too low, or too high? Was the approach in setting the passpoint correct, or could it be improved?

The three benchmark answers for each of the twenty situational interview questions are based on the real job behaviors collected by the CIT job analysis. The minimal 60% pass point indicates that the SMEs believe that overall a MAC should be able to provide the average answers for all the questions (i.e., 3*20=60).

Was 60% really the minimally acceptable pass point? After the examination was given, we reevaluated the way we applied the Angoff method and analyzed the data again. There were some concerns, as well as some fresh ideas.

Consider that both multiple choice test (MCT) and situational interview (SI) formats are similar because both are quantifiable, structured, and have multiple questions; yet the traditional MCT and SI is also different in that the MCT allows one and only one best choice; while different scores can be assigned on the SI depending on the degree to which a candidate answers the SI question correctly (e.g., either 5, 3, or 1 in the present study).

For example, when the SMEs are asked to estimate what percent of MACs would be able to answer a question correctly, only the best (5) answer was looked at; both the (3) and (1) answers where ignored. We may ask, if 60% of the MACs would be able to respond with the (5) answer, what about the other 40% of MACs? Would they all have missed the question? No, the other 40% of all of the MACs would likely give a (3) or (1) type of response. Of course, some of them would have missed the question entirely.

We are therefore proposing a more rational and precise MAC judgmental process as applied to SIs by the following: Ask the SMEs to distribute 100% among all possible choices (i.e., distribute among 5, 3, 1, & 0 answers). For each question, the weighted sum of the scores is the estimated probability a MAC would be able to answer that question "right".

Assume that all the SMEs happen to assign 60% to the (5) answer, the estimated probability for MACs to answer that question "right" for that particular question could range from 60% to 84%.

Pass Point Setting Procedures - Stage 2

Twelve months later we invited the same SMEs to return and apply the suggested modified judgmental pass point setting procedure. Four of the seven SMEs returned and again were instructed in the basic steps commonly followed in the use of the Angoff method, as noted earlier in this paper (Livingston, and Zieky, 1982). The suggested modification was introduced within the training of the Angoff method. The SMEs were instructed to distribute a total of 100 percent over all situational question response options (i.e. distribute among #5, #3, #1, and 0 responses).

After the completion of the SMEs' judgments, the estimated minimum pass point for both forms A and B were derived by the following combination of ratings:

1. factor weighting of each response option,(#5 by 1.0, #3 by 0.6, #1 by 0.2, and an 0 type of response by 0.0)
2. sum each factored response type per item, and
3. compute the average of the item sums.

Results

The averaging of the item sums resulted in an estimated pass point of 78.64 for form A, and 78.60 form form B. The coefficient alpha internal consistency reliability estimate for both form A and B were .81 and .66.

Discussion

The pass points derived from the Stage 2 procedure, if used approximately only 70% compared with more than 90% of the Stage 1 procedure would have passed the SI. If the SI was the first or only test part in the examination procedure, this lower pass rate would indicate greater utility of the pass point than this papers initial attempt. But more evidently, the modified adaptation of the Angoff passing point method does present a more rational method in the setting of the SI pass point.

Conclusion

As testing professionals, there are two principal sets of guidelines that lead us in the process of pass point setting: Standards for Educational and Psychological Testing (American Educational Research Association, 1985) and the Principles for the Validation and Use of Personnel Selection Procedure (Society for Industrial and Organizational Psychology, 1987). Neither one of these documents specifically discusses how to set pass points. However, both of them indicate the kind of information, such as the rationale to be used, which should be included in the documentation of the pass point setting process.

There is no one best way to set a pass point for a test. It is most important to have sound, defensible rationale behind every pass point decision. This study showed the possibility of using the less subjective judgmental Angoff method to set the pass point for the Interview, more specifically, the Situational Interview.

In conclusion, we would like to repeat what a very wise person has said which has been quoted many times before and will be many times more: "Anyone who expects to discover the 'real' passing score .... is doomed to disappointment, for a 'real' passing point does not exist to be discovered. All any examining authority ... can hope for ... is that the basis for defining the passing score be defined clearly, and that the definition be as rational as possible." (Ebel, 1972, p.496).

References

American Educational Research Association (1986). Standards for Educational and Psychological Testing. Washington, D.C. American Psychological Association.

Angoff, W.H. (1971) Scales, Norms, and Equivalent Scores. In Thorndike(ed.), Educational Measurements, Washington D.C., American Council on Education, pp. 514-515.

Berk, R.A. (1986) A Consumer's Guide to Setting Performance Standards on Criterion Reference Tests. Review of Educational Research, 56, 1, 137-172.

Cascio, W.F., Alexander,R.A., and Barrett,G.V. (1988) Setting Cutoff Scores: Legal, Psychometeric, and Personnel Issues and Guidelines. Personnel Psychology, 41, 1-24.

Ebel, R.L. (1972) Essentials of Educational Measurement. Englewood, N.J. Presentice-Hall.

Flanagen, J.C. (1954) The Critical Incident technique. Psychological Bulletin, 51, 327-358.

Glass, G.V. (1978) Standards and Criteria. Journal of Educational Measurement, 15(4), 237-261.

Latham, G.P., and 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., and Saari, L.M.,Pursell,E.D., and Campion,M.A. (1980) The Situational Interview. Journal of Applied Psychology, 65, 422-427.

Lin, T.R. (1988) The Situational Interview Experience: A Large Scale Application on the Selection of School Custodians. Presented at the Advanced Selection Techniques Conference, Personnel Testing Council of Northern California, March 11, 1988, Sacramento, CA.

Livingston, S.A. and Zeiky, M.J. (1982) Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Educational Testing Services, pp. 15-29.

McClung, G.G. (1974) Considerations in Developing Test Passing Point. International Personnel Management Asscciation.

Nedelsky, L. (1954) Absolute Grading Standards for Objective tests. Educational and Psychological Measurement. 14(1),pp. 3-19.

Society For Industrial and Organizational Psychology,Inc. (1987). Principles for the Validation and Use of Personnel Selection Procedures. (3rd Ed.) College Park Maryland.:Author.

USING THE SOCIAL SKILLS INVENTORY IN PERSONNEL ASSESSMENT

Ronald E. Riggio
California State University at Fullerton

The Social Skills Inventory (SSI) is a 90-item (in its revised form), self-report measure of basic social skills. The inventory includes separate measures of several basic skill dimensions. Total score on the SSI reflects a global level of social skills--what might be termed social competence or social intelligence.

The Basic SSI Dimensions

Emotional Expressivity (EE) is skill in nonverbal sending, dominated by skill in sending emotional messages, but also including the nonverbal expression of attitudes, expression of dominance, and sending of cues of interpersonal orientation. Persons highly skilled in emotional expressivity are animated and "emotionally charged."

Sample EE items:
- Quite often I tend to be the "life of the party."
- I have been told that I have "expressive" eyes.
- When I get depressed, I tend to bring down those around me.

Emotional Sensitivity (ES) is skill in receiving and decoding the nonverbal and emotional communications of others. Emotionally sensitive individuals attend to the emotional cues of others, and are skilled in rapidly and correctly interpreting subtle cues of emotion.

Sample ES items:
- It is nearly impossible for people to hide their true feelings from me.
- People often tell me that I am a sensitive and understanding person.
- At parties I can instantly tell when someone is interested in me.

Emotional Control (EC) is the ability to control and regulate emotional and nonverbal displays. Emotional control includes ability to pose emotions on cue and ability to cover felt emotions with a posed emotional "mask." In extreme, the person very high on EC may tend to control the display of felt emotional states.

Sample EC items:
- I am able to conceal my true feelings from just about anyone.
- I am very good at maintaining a calm exterior, even when upset.

Social Expressivity (SE) is skill in verbal expression and the ability to engage others in social discourse. High scores on the scale of Social Expressivity are associated with verbal fluency, ability in initiating conversations, and ability to speak spontaneously on a topic.

Sample SE items:
- When in discussions, I find myself doing a large share of the talking.
- I usually take the initiative and introduce myself to strangers.

Social Sensitivity (SS) refers to verbal receiving ability and a sensitivity to, and understanding of, the norms governing appropriate social behavior. Socially sensitive persons are attentive to social behavior and conscious and aware of the appropriateness of their own actions.

Sample SS items: • I often worry that people will misinterpret something that I have said to them.

• While growing up, my parents were always stressing the importance of good manners.

Social Control (SC) is skill in role-playing and social self-presentation. Persons high in the skill of social control are socially adept, tactful, and socially self-confident. They have an ability to fit in to just about any type of social situation.

Sample SC items: • I find it very easy to play different roles at different times.

• When in a group of friends, I am often spokesperson for the group.

• I can fit in with all types of people, young and old, rich and poor.

Table 1: Correlations Between Total Score on the SSI and Social Behaviors & Personality Dimensions

Self-Reported Social Behaviors	SSI	Personality Dimensions	SSI
Acting Experience (n=60)	.22*	Extraversion (Eysenck) (n=85)	.08
Sales Experience (60)	.25*	Self-Monitoring Scale (149)	.53***
Number of Close Friends (59)	.49***	Affective Communic. Test (149)	.78***
Number of Acquaintances (57)	.40***	Social Desirability Scale (149)	.04
Public Speaking Comfort (60)	.36***	Social Anxiety Scale (149)	-.52***
Shyness	-.53***	Social Support Scale (127)	.24**

* $p < .10$; ** $p < .05$; *** $p < .01$

References

Riggio, R.E. (1986). Assessment of basic social skills. Journal of Personality and Social Psychology, 51, 649-660.

Riggio, R.E. (1987). The charisma quotient. New York: Dodd, Mead.

Riggio, R.E. (in press). Manual for the Social Skills Inventory. Palo Alto, CA: Consulting Psychologists Press.

Riggio, R.E. (in press). Social skills and interpersonal relationships: Influences on social support and support-seeking. In W.H.Jones & D. Perlman. (Eds.) Advances in interpersonal relationships. Vol. 2, Greenwich, CT: JAI Press.

THE EFFECT OF PAQ ITEM TYPE ON ANALYST INTERRATER RELIABILITY

Calvin C. Hoffman
Southern California Gas Company,

Lisa M. Holden
California State University - Long Beach,

and Jade Hoffman
Los Angeles Unified School District

ABSTRACT

This study examined the level of interrater reliability found for four categories of PAQ items. The 190 PAQ items were sorted into the following categories: (1) Special code (S-code) items, (2) anchored items (anchors refer to average ratings for benchmark jobs), (3) non-anchored items, and (4) factual items. A total of 24 jobs were analyzed using three analysts each (72 PAQ'S).

Results indicated that S-coded items are rated more reliably than the complete PAQ. Anchored items were rated less reliably than were non-anchored items, which was probably a function of the large number of Does Not Apply (DNA) ratings for the non-anchored items. Factual items were rated more reliably than were S-coded items. The results have implications for the training of PAQ analysts.

INTRODUCTION

Previous research on the PAQ has examined the effects of variables such as providing job analysts with less information prior to making ratings (Jones, Main, Butler, & Johnson, 1982), or varying the level of rater expertise (Cornelius, De Nisi, & Glencoe, 1984). Both studies found relatively low levels of interrater reliability; Jones et al (1982) reported a median interrater correlation of .48.

Harvey & Hayes (1988) demonstrated that high frequencies of Does Not Apply (DNA) ratings on the PAQ could mask substantial rater disagreement on the remaining elements which do apply to a job. Other questions can be raised about the rating task which the PAQ poses to job analysts.

Based on our use of the PAQ, certain items, and in particular, certain types of items, are much easier to rate than others. The 190 PAQ items were independently sorted by the first two authors into four separate groups. The four groups are as follows: (1) 21 Special coded (S-code) items, (2) 66 anchored items (anchors refer to average ratings for benchmark jobs), (3) 81 non-anchored items, and (4) 22 factual items. (Four of the 194 PAQ items are blank, so the total item pool was 190 items.)

HYPOTHESES

1. S-code items will be rated more reliably than either anchored or non-anchored items.
2. Anchored items will be rated more reliably than non-anchored items.
3. Factual items will be rated more reliably than non-anchored items.

METHOD

ANALYSTS AND JOBS

A total of six analysts were involved in analyzing 24 jobs. Each job was analyzed by three analysts. All analysts received one and a half days of training prior to rating the jobs. Due to scheduling constraints, various combinations of analysts rated the jobs.

RESULTS

Across all PAQ's, the average interrater reliability was $\underline{r} = .74$. S-Code ratings were rated more reliably with an average $\underline{r}$ of .78. Contrary to expectations, non-anchored items were rated more reliably (average $\underline{r} = .70$) than were anchored items (average $\underline{r} = .66$). Finally, factual items were rated at very high levels of reliability (average $\underline{r} = .84$). In several cases, average reliability for factual items on a specific job, was 1.00.

A multiple regression was performed, treating average full-scale reliability as the criterion, and average reliability on each of the four item categories as predictors. Total scale reliability was predicted quite well by the four item categories ($\underline{R} = .976$, $\underline{p} < .0001$) (see Table 1). Examination of the regression weights for the item categories reveals that relative weight of each category parallels the relative frequency of items in the category, and hence, number of N ratings. A notable exception is the factual item cateogry. Even though this category had the highest average reliability, and the highest percentage of N ratings, it was not predictive of full scale PAQ reliability.

TABLE 1

REGRESSION ANALYSIS: PREDICTION OF FULL SCALE PAQ RELIABILITY WITH AVERAGE RELIABILITIES OF FOUR ITEM CATEGORIES

MEASURES	PARAMETER ESTIMATES	STANDARD ERROR	t	p
Intercept	.1342	.0361	3.718	.0015
S-Code	.1696	.0397	4.267	.0004
Anchored	.2193	.0309	7.089	.0001
Non-Anchored	.4734	.0600	7.896	.0001
Factual	-.0051	.0247	-0.208	.8371

$R^2 = .952$

$\underline{F}\ (4,19) = 95.019$ $\underline{p} < .0001$

DISCUSSION

These results demonstrate that some categories of PAQ items are rated much more reliably than others. Contrary to expectations, anchored PAQ items were rated less reliably than were non-anchored PAQ items. Since it is clear that high percentages of N ratings can help increase apparent reliability (Harvey & Hayes, 1986), and since the non-anchored items have a much higher percentage of N ratings, it is not clear to what extent the use of anchors affects rating reliability on the PAQ. Clearly, the S-code items were rated more reliably than was the complete PAQ; S-code items are also much better defined than are other items in the PAQ.

Based on the results of the study, one might suggest that a higher percentage of PAQ items be defined as are the S-code items. Given the clear definitions in both the PAQ and the job analysis manual, providing such definitions would make the rating task considerably easier, and hence should increase interrater reliability. Conversely, the fact that anchored items were rated less reliably than non-anchored items could suggest that anchoring more PAQ items might not necessarily result in increased interrater reliability.

Training of PAQ analysts could be altered to emphasize anchored and non-anchored items in terms of defining the element being evaluated. Relatively less time could be spent on S-code items, since the elements are so well-defined and more self-explanatory. Likewise, little time need be spent on the factual items. Information to rate those items should be readily available to analysts, and rater reliability should not be a problem. It would be useful to replicate this study with a larger sample of jobs covering a wider range of occupations. This would help clarify whether the results of this study are due to the nature of the jobs analyzed here, or to differences in the way PAQ elements are defined and/or anchored. Finally, one should recognize that analyst experience and training will affect the results of any examination of interrater reliability of ratings.

A complete version of this paper is available on request to the first author. Address: Southern California Gas Company, 810 South Flower Street, Mail Location 303H, Los Angeles, CA 90017

* * * * * * * * * * *

EMPLOYEE APPEALS IN THE FEDERAL SECTOR

Paul van Rijn
U.S. Merit Systems Protection Board

This poster session highlighted the annual reporting by the U.S. Merit Systems Protection Board (MSPB) of the number and types of appeals it decided during fiscal year (FY) 1987. MSPB is the quasi-judicial Federal agency charged, in part, by Congress to adjudicate appeals from Federal employees, annuitants, and applicants concerning certain personnel actions taken by Federal agencies.

MSPB was created by the Civil Service Reform Act (1978) and charged to continue the mandates of the Pendleton Act (1883) to protect the integrity of Federal merit systems against prohibited personnel practices, to ensure adequate protection for employees against abuses by agency management, and to require Federal executive agencies to make employment decisions based on individual merit.

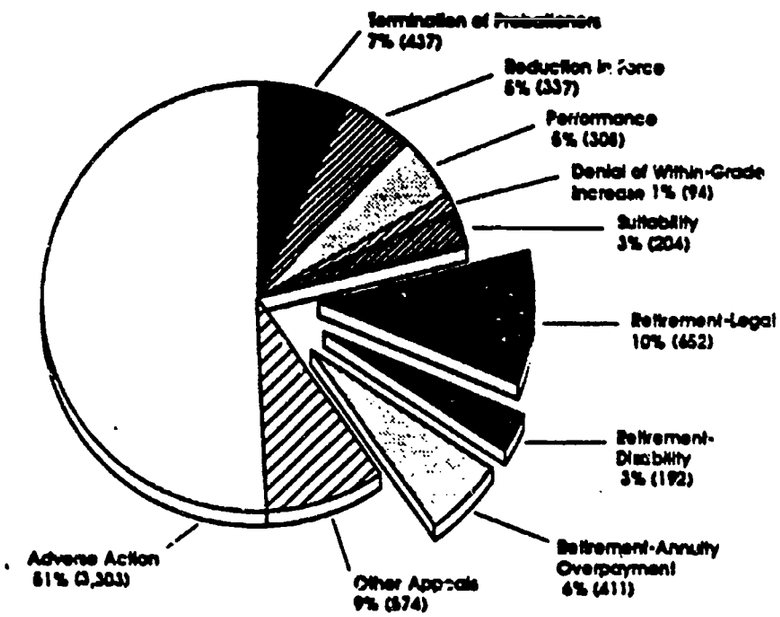
Over 6,500 initial appeals were decided by MSPB's administrative judges during FY 1987. Thirty-nine percent of these appeals were dismissed because they were not filed in a timely manner or were outside MSPB's jurisdiction. Of the cases not dismissed, 36 percent were settled by mutual agreement between the parties. This rate of settlement represents a substantial increase over the 26-percent settlement rate of FY 1985 and the 6 percent rate in FY 1984, when settlement data were first reported.

Figure 1 shows that most (51 percent) of the initial appeals were based on adverse actions (e.g., reduction in grade, suspension, furlough) by the agency. Nineteen percent addressed disagreements over retirement issues, while the remaining appeals were based on termination of probationers, reductions in force, removals for inadequate performance, suitability determinations, and other appealable agency actions.

Of the 2,540 initial appeals that were not dismissed or settled, 76 percent affirmed the agency action. Twenty-four percent reversed, modified, mitigated, or otherwise changed the agency action.

Twenty-one percent of the initial appeals included allegations of discrimination. Of these appeals, discrimination was found in only 4 percent of the cases. The most frequent allegation of discrimination was handicapping condition (39 percent), followed by race (25 percent).

Figure 1. Types of Initial Appeals Decided in FY 1987



Over half (54 percent) of the initial appeals were presented by persons with little or no legal training (e.g., the appellant, a friend, a co-worker).

Decisions by MSPB administrative judges become final unless the decisions are petitioned for a (second-level) review by the three-member bipartisan Board of MSPB. The Board issued final decisions on 1,619 petitions for review during FY 1987.

In addition to the initial appeals and petitions for review, MSPB issued decisions and conducted activity in a variety of other cases and adjudicative matters. Board decisions were affirmed by the U.S. Court of Appeals for the Federal Circuit in 99 percent of the 375 cases that it adjudicated on the "merits" of the case during FY 1987.

The report on which this poster session was based is entitled, A Study of Cases Decided by the U.S. Merit Systems Protection Board in Fiscal Year 1987, and may be obtained from the author at the U.S. Merit Systems Protection Board, 1120 Vermont Avenue NW, Washington DC, 20419.

* * * * * * * * * * *

THE WORKER CHARACTERISTICS INVENTORY: A METHODOLOGY FOR ASSESSING PERSONALITY DURING JOB ANALYSIS

Steven Arneson
Hogan Assessment Systems, Inc.

It has long been accepted that a complete job analysis should contain data regarding the work itself (tasks, activities, equipment, etc.) as well as information about the worker (education, experience, KSA's, etc.). Job analyses that address both of these issues are more than adequate for describing the nature of the job and the minimal qualifications for performance. But herein lies the problem: aside from this information, traditional job analysis procedures have done little in the way of identifying characteristics of effective or successful workers. Job analysis should extend beyond merely defining tasks and minimal qualifications, it should also be used to describe what type of person will do well in a particular job. Unfortunately, current job analysis procedures

generally do not include a systematic method for identifying these characteristics. Research needs point to a measurement device that describes the personal attributes necessary for successful performance. Such an instrument should be grounded in personality theory, and it should be short and easy to use. This paper introduces an instrument designed to meet these requirements.

The Worker Characteristics Inventory (WCI) is a theory-based personality checklist designed specifically for use during job analysis. The WCI consists of 80 true-false adjective items; these items form the content for six personality scales associated with social and occupational success. These scales are: Intellectance, the degree to which a worker is seen as intelligent, well-educated, and interested in ideas; Adjustment, the degree to which a worker seems free of the everyday symptoms of maladjustment; Prudence, the degree to which a worker seems dependable, conscientious, and reliable; Ambition, the degree to which a worker seems hard-working, energetic, and leaderlike; Sociability, the degree to which a worker is gregarious, affiliative, and outgoing; and Likeability, the degree to which a worker seems agreeable and pleasant. The WCI is included as part of the job analysis questionnaire, and incumbents and/or supervisors respond by identifying the personality characteristics of the ideal worker for that particular job.

To identify the personality characteristics of effective employees, the WCI was administered to 735 incumbents and 85 supervisors in 13 occupational groups. Four research hypotheses were examined. First, of primary concern is whether the WCI distinguishes between jobs. If only one personality profile exists for all workers, the WCI will not have much utility for individual occupational groups. However, in all likelihood, certain personality traits exist in varying degrees of importance for different jobs. The first research question then, may be stated as follows: will the WCI produce distinct profiles for workers in different occupations? To answer this question, WCI profiles for the seven largest subgroups of the research sample were examined. Twenty-one individual two-group discriminant analyses were performed to assess the difference between

each of the possible group pairs among these seven groups. Of these 21 discriminant analyses, 18 resulted in significant differences between occupational groups ($p < .001$).

Second, job analysis users need to be concerned about the degree to which incumbents and supervisors agree about the ideal worker characteristics. This question is useful for determining the appropriate response group to use the WCI. If in fact individuals use similar trait vocabularies to describe others, and if incumbents and supervisors have the same view of what personal qualities are necessary to perform effectively, then the two rater groups should generate similar profiles. The second research question may be stated as follows: will job incumbents and supervisors differ in their description of the ideal worker characteristics using the WCI? WCI ratings from incumbents and supervisors were available for 9 occupational groups. Discriminant analysis results revealed no significant differences between incumbent and supervisor WCI profiles for eight of the nine groups.

Third, is actual job experience necessary to profile accurately the ideal worker characteristics? This question has important implications for users who may want job analysts to complete the WCI. To study this question, a group of naive raters (N=[illegible]) were asked to read brief job descriptions of four jobs and describe the ideal worker for this job by completing the WCI. These results were then compared with the WCI profiles generated by incumbents. Thus, the research question is: will non-incumbents and actual employees differ in their description of the ideal worker characteristics using the WCI? Results did not support the hypothesis that naive raters and job incumbents would produce similar ideal worker profiles. Four separate discriminant analysis procedures were performed to determine the difference between naive raters and job incumbent profiles. Significant differences ($p < .001$) were found for all four occupational groups.

Finally, there is the question of whether the WCI simply identifies the traits that are characteristic of a "good person", regardless of the incumbent or job in question. This issue is important for determining whether or not the

WCI is simply measuring halo (describing good workers in all jobs) or whether it is actually providing descriptive statements of ideal worker characteristics. The question is: will profiles of the "ideal worker" differ from profiles of the "ideal person"? Four discriminant analyses were performed to determine the difference between naive rater profiles of the ideal person and these same raters' profiles of ideal workers. These results reveal significant differences between profiles for each of the four occupational groups ($p < .001$).

Assessing personality in organizations is not a novel concept; industry has used personality assessment in one form or another since the turn of the century. Indeed, people engage in impromptu assessments of others every day, using trait terms to describe consistencies in interpersonal behavior. When asked, people can think of dozens of trait terms to describe co-workers (hard-working, lazy, reliable, consistent, careful, cooperative, etc.).

Results of this study indicate that the Worker Characteristics Inventory is a reliable technique for identifying the personal qualities that describe successful workers in a particular job. Because this information has utility for a number of personnel related decisions, the WCI has been designed specifically for use as a component of job analysis. For years, job analysis experts have been advocating the collection of job information that details the "other personal characteristics" contributing to job effectiveness. Now, in addition to collecting data about tasks and KSA's, job analysts may use the Worker Characteristics Inventory to assess systematically the personality traits that describe the ideal worker.

* * * * * * * * * * *

REFINEMENT OF A SELF-RATING SELECTION INSTRUMENT: CORRECTION OF SELF-BIAS[1]

Walter G. Mann, Jr.
U.S. Office of Personnel Management
Washington, D.C.

Compared to written tests of maximum performance, self-ratings are inexpensive, readily accepted, easily administered, and less subject to adverse impact. Their main drawback is fakability (Levine, Flory, & Ash, 1977; van Rijn, 1980; and Mabe and West, 1982).

A valid self-appraisal instrument, corrected for self-bias, would be a very promising selection device. Attempts have been made to measure self-bias--most notably, Anderson, Warner, & Spencer (1984)--but to date no one has been able to show that their measure of self-bias is anything other than general cognitive ability. For example, what Anderson and associates called self-bias seems to be nothing more than a lack of knowledge of English phrases, and might more appropriately be called verbal ability.

My introduction to self-bias came in 1967 as a result of analyzing some job element data on about a thousand clerical applicants. Factoring the intercorrelations between 6 test scores and 10 self-ratings, I found three factors: Verbal, Quantitative/Clerical Speed, and a third factor. On the third factor the self-ratings of job elements loaded positively and test scores loaded negatively. I tentatively named the third factor Self-Bias, but did not have enough support for publication.

A few years ago I decided to test the feasibility of using self-ratings as a criterion for the validation of a selection test (Mann, 1984). At that time OPM was validating tests in as many ways as possible. Realizing that we might not always have such resources, I decided to study an inexpensive and quick method of validation using applicant's self-ratings as the criterion. After I correlated the test scores with the self-rating criterion, I put away the results to check them later against the results for conventional criteria. When I heard that OPM was developing a biographical instrument to select professional and administrative career (PAC) employees, I decided to use the data from the apprentice validation study to do some exploratory research on self-bias. The objective was not to develop a self-bias measure that could be used to correct self-appraisals in the PAC biographical instrument, but merely to provide some exploratory research on the development of a measure of self-bias, with the hope that it might be of some small assistance to the PAC researchers.

[1]The author has additional results which could not, because of space limitations, be included in the present paper.

METHOD

The subjects were 2,593 applicants for apprentice trade and craft positions with a federal agency in a large, southeastern city. The subjects were administered a battery of ability tests. They were asked, on a voluntary basis, to provide self-ratings on 19 knowledges, skills, abilities, and other characteristics. Responses were made on a four-point scale. Subjects were able to indicate if they could not rate themselves on a characteristic.

I decided, mostly for purposes of multiple regression analysis, that I wanted to have complete data on all subjects. Therefore I dropped from the study those who had not rated themselves on all 19 characteristics. This left me with 2,119 cases. Because of the large N and the exploratory nature of the research I decided against sophisticated analyses, such as double-hold-out samples.

I grouped the 19 self-rating characteristics based on high intercorrelations with one another. Eight groups resulted: quantitative reasoning (QR), verbal, perceptual, following directions, short-term memory, perceptual speed, psychomotor, and overall. I grouped characteristics to make the reliability of the self-ratings more comparable to the reliability of the ability tests in the study, to cut down on unnecessary redundancy, and also to have approximately the same number of self-rating scores and test scores in the factor analysis. Several tests were not used in the factor analysis because they correlated over .50 with another test. For each individual, for each of the eight characteristics, a simple sum of ratings of appropriate characteristics was used as an estimate of ability. For example, a self-rating estimate of QR was obtained by adding the self-ratings of three QR characteristics.

Two factor analyses were run: the first, to assist in the interpretation of a self-bias factor; a second, to generate factor loadings that would be used to compute self-bias factor scores. For purposes of the first factor analysis, an experimental measure of self-bias was obtained by subtracting, for each individual, the standardized test score for QR from the standardized sum of the self-ratings for QR. This measure of self-bias is based on estimates of quantitative ability and therefore cannot legitimately be used to predict a criterion measure of QR. An appropriate measure of self-bias for predicting QR must be experimentally independent of QR. The scores for all the tests (except QR) and self-ratings (except QR) were intercorrelated, and the results factor analyzed and rotated obliquely. A priori, a self-bias factor was defined as one which has positive loadings on self-ratings and negative loadings on tests (or vice versa, since factor loading signs can be changed without changing the meaning of a factor.) In addition, the experimental measure of self-bias should load positively on the self-bias factor.

The second factor analysis was the same as the first, except that the experimental measure of self-bias was omitted. The factor loadings of the second factor were used to generate Self-Bias factor scores for each subject.

A multiple regression was run, using the self-rating estimates of QR and the Self-Bias factor scores as the predictors, and QR test scores as the criterion measure.

RESULTS AND DISCUSSION

The factor analysis used for interpretation yielded two factors. All variables loaded positively on the first factor and was named G (for general ability). The second factor fulfilled the a priori requirements for self-bias. All the self-rating characteristics loaded positively on this factor, while all the tests loaded negatively; the experimental measure of self-bias loaded .49, the highest positive loading of any variable.

TABLE 1

Promax Factor Loadings for Two Factors
(N=2,119)

	Factor	
	G	Self-Bias
Experimental Self-Bias Measure	.27	.49
Self-Rating Characteristics		
Verbal	.59	.17
Follow Directions	.67	.26
Short-Term Memmory	.58	.25
Perceptual Speed	.71	.20
Perceptual	.59	.20
Psycho-motor	.65	.38
Overall	.63	.22
Tests		
Measuring	.28	-.41
Perception	.24	-.36
Spelling	.26	-.44
Oral Directions	.25	-.47
Arithmetic	.29	-.48
Eye-Hand Coordination	.27	-.38
Table Reading	.34	-.59

The measures of Self-Bias in the present study were not strongly related to G. The correlation between the Self-Bias factor and the G factor was only .18. In addition, the experimental self-bias measure loaded only .27 on the G factor.

The self-rating estimate of QR was a reasonably good predictor of QR test scores (r = .51). More importantly, the Self-Bias factor scores added significantly to the multiple correlation (R = .64). The Self-Bias factor scores were able to add unique predictor variance because of their correlation with the criterion measure (r = -.34); i.e. they did not operate as a suppressor variable. The Self-Bias factor scores correlated .09 with the self-rating estimate of QR.

Applicant data presumably produced more inflation in self-ratings than incumbent data would have, and this might helped in the measurement of self-bias. On the other hand, use of applicants meant there was no measure of job performance. Ergo the decision to use a test as the criterion. Obviously, a study needs to be done in which job performance is the criterion.

The Self-Bias factor scores were based, in part, on test scores. If test scores had not been used, the validity of the Self-Bias factor scores would have dropped significantly. This should not be a problem in a PAC examination because the biographical instrument would be used in conjunction with tests of maximum ability.

The presence of a self-bias measure could indirectly prove useful if it discouraged applicants from giving inflated self-ratings.

It should be obvious that self-bias is not a well-defined construct. This should not preclude its use for selection because at present we have only a few well-defined psychometric constructs for use in hiring.

For those inclined to work with self-bias, I would recommend including in the research plan other measures of self-bias, such as honesty or lie scales found in some personality inventories, with the intention of developing a nomological network.

CONCLUSIONS

It is possible to develop factor scores, for what tentatively has been named Self-Bias, that predict an objective measure of performance, quantitative reasoning test scores. Self-ratings of quantitative reasoning also have validity for predicting quantitative reasoning test scores. In combination the Self-Bias factor scores and the self-ratings have even greater validity.

REFERENCES

Anderson, C.D., Warner, J.L., & Spencer, C.C. (1984). Inflation bias in self-assessment examination: Implication for valid employee selection. Journal of Applied Psychology, 69, 574-580.

Levine, E.L., Flory, A., & Ash, R.A. (1977). Self-assessment in personnel selection. Journal of Applied Psychology, 62, 428-435.

Mabe, P.A. and West, S.G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.

Mann, W.G. (1984). Correction of Ratings of Ability for Self-Bias. Presented at the annual meeting of the American Psychological Association, Toronto, Canada.

van Rijn, P. (1980). Self-assessment for personnel examining: An overview (PRR-80-14). Washington, D.C.: U.S. Office of Personnel Management.

* * * * * * * * * * *

THE SITUATIONAL INTERVIEW VERSUS SELF-ASSESSMENT: WHAT CAN BE DONE IF CANDIDATES INFLATE THEIR SCORES?

Carol L. Manligas & Thung-Rung Lin
Los Angeles Unified School District

The main purpose of this study was to test the hypothesis that job candidates would not inflate their scores on a situational interview (SI) or mixed-standard scale self-assessment checklist (MSSSAC).

Latham and his associates (Campion, Pursell, & Brown, 1988; Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980) introduced the situational interview (SI) method. This SI format is highly structured and content valid. From a critical incident job analysis data base, interview questions are developed with corresponding benchmark responses. The interview questions describe situations which current job incumbents encounter on the job. The benchmarks represent different levels of performance and are assigned appropriate values. Candidates respond by indicating what they would do in the situation described and receive the score that represents their response relative to the benchmarks. Two recent studies have reported satisfactory predictive validities ranging from .45 to .56 (Campion, et al., 1988; and Weekley and Gier, 1987).

Self-assessment on the other hand, as the term suggests, refers to the estimates of achievements or capabilities which job applicants make of themselves. Previous studies have characterized self-assessment devices as being: (1) high in inflation bias, (2) not reliable, and (3) lacking in discriminability which reduces its predictive validity in the employment selection process (Anderson, Warner, & Spencer, 1984; Mabe and West, 1982; van Rijn, 1980; Levine, Flory, & Ash, 1977). However, if the predictive validity is improved, self-assessment could be the most cost-effective selection method in employment testing especially when the candidate population is very large.

In the performance evaluation area, researchers have attempted to improve self-assessment in job performance evaluation by incorporating the mixed-standard methodology (Blanz & Ghiselli, 1972), which reduces transparency by eliminating the recognition of order-of-merit in the behavioral dimensions being rated. In the area of employment selection, Anderson, et al. (1984) used two different methods to reduce inflation bias: embedding a lie scale and statistically adjusting scores.

The application of the mixed-standard scale methodology to self-assessment in personnel selection was first introduced by Lin, Magel, and Manligas (1986). They created a situational interview (SI) and mixed-standard scale self-assessment checklist (MSSSAC) for personnel selection from the same critical incident job analysis (Flanagan, 1954) data base. Their results indicated that the MSSSAC yielded a more normal distribution of scores which contradicted the conventional belief that self-assessment in the employment setting is always inflated (i.e., skewed towards the positive side) and lacked discriminability. Although the overall correlation between SI and MSSSAC was not significant, comparing the MSSSAC with SI, two job factors out of five, i.e., safety awareness and initiative, positively correlated with each other ($r=.31$ and .23, $p<.01$).

In a follow-up study, Lin and Manligas (1987) reported a six-month test-retest reliability estimate of .80 on MSSSAC based on a sample of 35 School Custodial incumbents. This contradicted the general belief that self-assessment is not a reliable measure. In the same study, they also compared the MSSSAC with a simple self-assessment checklist (SAC), however; no relationships were found between the MSSSAC and SAC. They attributed the failure to find relationship between MSSSAC and SAC to a highly inflated and very negatively skewed distribution on SAC scores.

The purpose of this study is to replicate our 1986 study by using actual School Custodial candidates instead of job incumbents and to assess the robustness of the SI and MSSSAC in personnel selection by incorporating an inflation scale (Anderson, et al., 1984).

Hypothesis 1: Inflation will have no effect on either the MSSSAC or SI.

Hypothesis 2: The MSSSAC scores will validly predict the SI scores, providing that they are both based on the same job analysis data base and assessing the same job factors.

METHOD

A sample of candidates from the 1987 examination for positions of School Custodians for a west coast urban school district (N = 284) served as subjects. These candidates were asked at the end of the examination to voluntarily complete a questionnaire. Both the SI and MSSSAC were constructed from the same critical incident job analysis data base. The SI was used as the final hurdle for a multiple hurdles examination which included a willingness to work checklist, written test, and a reference check. For a more detailed explanation of the development and procedures of both SI and MSSSAC, please refer to Lin (1988) and Lin, et al. (1986).

Two different MSSSAC forms were used (i.e., with/without inflation scale) because we were also interested in knowing the impact of the inflation scale on self-assessment. For the scoring of the MSSSAC, the statements that corresponded to the same SI question were used together to determine a value for that particular item. These scoring combinations are in agreement with the rationale used for the scoring of SI questions.

In order to make a comparison to previously published self-assessment studies, such as Anderson, et al. (1984), an inflation scale was created for the MSSSAC. Five implausible behavioral statements that, on the surface, appeared to be similar to the real MSSSAC items were created. They represented impossible custodial job behaviors. Subjects were rated on these five implausible behavioral statements using the same scales as in MSSSAC.

For comparison with Anderson, et al. (1984), both regression fomula and inflation proportion methods were used to correct for inflation bias.

RESULTS

Reliability estimates were calculated for all three scales used. The reliabilities for the two SI forms are .71 and .66. For the two MSSSAC forms, they are .80 and .65. The five implausible custodial behavioral statements are designed to measure the inflation bias, which yield an internal consistency reliability of .64. An analysis of the IP values received by subjects in this study suggests that the attempt to inflate on the MSSSAC was extensive.

Robustness of MSSSAC and SI. Although not hypothesized, we also tested whether or not introducing the "inflation scale" itself would have an impact on the MSSSAC scores. No significant difference was found between these two groups. One purpose of this study was to test the hypothesis that inflation scale scores will have no effect on either SI or MSSSAC. To test this hypothesis, the inflation scores were correlated with MSSSAC and SI. The pearson-product correlation between the MSSSAC and inflation scale scores was -.27 ($p<.001$, N=173), while no significant relationship was found between the SI and inflation scale scores. Both the MSSSAC and SI were based on the same job analysis data base and covered essentially the same job behaviors. The Pearson-product correlation between these two measures was .28 ($p<.001$, N=201).

In order to assess whether statistical correction for the inflation scale scores affects the predictability of the MSSSAC on the SI scores, a correlation between the SI and corrected MSSSAC scores (i.e., Xcm) was found at .33 ($p<.001$, N=201).

The age and sex of the subjects were not significantly related to the SI, MSSSAC, and Inflation scale scores. Ethnicity was also found to have no effects on Inflation scale scores. However, significant relationships were found between ethnicity and both the SI and MSSSAC scale scores. For more discussion of race effects on the SI scores, please see Lin and Manligas (1988).

DISCUSSION

This study is consistent with the literature as it demonstrates that inflation bias is prevalent in self-assessment when used in the context of personnel selection. Inflation bias was found even when candidates knew that the MSSSAC score would not be used in the selection decision.

Although we found that there was no effect on SI, we found a negative correlation between the inflation scale and the MSSSAC. The results indicate that the higher the inflation scale score, the lower the MSSSAC score. This is attributed to the robustness of the MSSSAC method. Perhaps, candidates who tried to exaggerate their scores would respond "I would do this." to most of the statements, including the items in the inflation scale. However, the scoring of the MSSSAC depends on the logical relationship within the related triad of statements.

We found the relationship between MSSSAC and SI scores moderately significant at .28. When the MSSSAC score is statistically adjusted for inflation bias (i.e., one of the methodologies proposed by Anderson, et al., 1984), the correlation increases from .28 to .33. This slight improvement implies that the MSSSAC method has an inherent mechanism that reduces inflation bias at an effective level. Using the MSSSAC as a valid and cost-effective self-assessment predictor in personnel selection appears promising.

Previous studies have shown that only certain abilities (e.g., typing ability) can validly be predicted by self-assessment devices in comparison to other selection methods. The significant correlation between the SI and MSSSAC scores is encouraging because both the SI and MSSSAC are global measures of a combination of job factors, e.g., attendance, job awareness, safety awareness, initiative, and interpersonal relations. By identifying specific job factors, we believe stronger correlations of certain job factors between the SI and the MSSSAC will emerge.

Future studies should look into the cognitive process which occurs when candidates answer the MSSSAC and how they interpret the scales (e.g., "I would do this differently."). Do they consider this in a positive or negative way? It is our recommendation that testing this methodology and replicating the study on other higher job classes which include certain KSAs, not easily measured by any test part, would further identify the usefulness of this MSSSAC/SI approach in reducing inflation bias in personnel selection.

REFERENCES

Anderson, C. D., Warner, J. L., and Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. Journal of Applied Psychology, 69, 574-580.

Blanz, F. and Ghiselli, E. E. (1972). The mixed standard scale: A rating system. Personnel Psychology, 25, 185-199.

Campion, M. A., Pursell, E. D., and Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment

interview. Personnel Psychology, 61, 25-42.
Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.
Latham, G. P. and Finnegan, B. J. (1987). The practicality of the situational interview. Unpublished Manuscript.
Latham, G. P. and Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. Journal of Applied Psychology, 69, 569-575.
Latham, G. P., Saari, L. M., Pursell, E. D., and Campion, M. A. (1980). The Situational Interview. Journal of Applied Psychology, 65, 422-427.
Levine, E. L., Flory, A., and Ash, R. A. (1977). Self-assessment in personnel selection. Journal of Applied Psychology, 62, 428-435.
Lin, T. R. (1988). The situational interview experience: A large scale application on the selection of school custodians. Presented at the Advanced Selection Techniques Conference, Personnel Testing Council of Northern California, March 11, 1988, Sacramento, CA.
Lin, T. R., Magel, S., and Manligas, C. (1986). A comparative analysis of the situational interview and self-assessment checklist in the employment selection process. Presented at the 10th Annual Conference of the International Personnel Management Association Assessment Council, San Francisco, CA June, 1986.
Lin, T. R. and Manligas, C. L. (1988). A comparative analysis of rater and ratee race effects in employment selection interviews: Structured interview versus situational interview. To be presented at the Annual Conference of the Academy of Management, Anaheim, CA, August 8-10, 1988.
Lin, T. R. and Manligas, C. L. (1987). A comparative analysis of two self-assessment techniques for Custodian. Presented at the 11th Annual Conference of the International Personnel Management Association Assessment Council, Philadelphia, PA, July, 2, 1987.
Mabe, P. A. and West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.
van Rijn, P. (1980). Self-assessment for personnel examining: An overview, (PRR-80-14) Washington, D. C.: U. S. Office of Personnel Management, Personnel Research & Development Center.
Weekley, J. A. and Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. Journal of Applied Psychology, 72, 484-487.

* * * * * * * * * *

COMPUTERIZED TESTING MADE PRACTICAL:

THE COMPUTERIZED ADAPTIVE EDITION

OF THE DIFFERENTIAL APTITUDE TESTS

James R. McBride

The Psychological Corporation
San Diego, California

Introduction

This is a brief description of the first commercially published test to employ the technology called "computerized adaptive testing", in which a computer is used as the test administration medium, and the difficulty of each test is tailored to the performance level of each examinee.

Since 1947, The Psychological Corporation has published the Differential Aptitude Tests, a battery of eight tests used for educational placement and for vocational guidance counseling, primarily in junior and senior high schools, and for personnel assessment and selection in business and industry.

The Differential Aptitude Tests (DAT) have been revised a number of times over the years, and are highly regarded for their usefulness and technical quality. The Computerized Adaptive Edition of the DAT was first published in 1986, for use on Apple // series microcomputers; a second version, published in 1988, operates on IBM PC and compatible personal computers.

The Adaptive DAT computer software is capable of administering all eight DAT subtests, and the optional Career Planning Questionnaire. It scores each test immediately, and is capable of providing immediate results to the user, either as scores displayed on the computer screen or as printed reports.

All seven power tests of the DAT are administered adaptively; as a consequence of adaptive administration, they are only half the length of their printed counterparts. The eighth test -- Clerical Speed and Accuracy -- is a highly speeded test; the computer times its administration, probably more accurately that it is timed in typical classroom testing with the printed edition.

The reduced length of the seven adaptive tests makes the Adaptive DAT considerably more efficient than the printed edition. The printed edition typically takes about three and a half hours to administer. In contrast, the Adaptive DAT typically takes less than two hours.

This short paper will describe the design of the adaptive edition, the calibration of DAT test items for use in the adaptive edition, and some of the empirical research that has been conducted to compare the Adaptive DAT with its printed counterpart.

The Differential Aptitude Tests

Some background on the DAT was presented in the introduction. This section will briefly give some additional information about the current edition of the printed DAT.

The DAT consists of eight tests, seven of which are essentially power tests, and one of which is a short speeded test. The tests' names and standard abbreviations are listed below:

VR	Verbal Reasoning	50 5-choice items
NA	Numerical Ability	40 5-choice items
AR	Abstract Reasoning	45 5-choice items
CSA	Clerical Speed and Accuracy	100 5-choice items
MR	Mechanical Reasoning	70 3-choice items
SR	Space Relations	60 4-choice items
SP	Spelling	90 2-choice items
LU	Language Usage	50 5-choice items

All seven of the power tests are timed; however, the time limits are fairly generous. Of the seven, the first five are primarily aptitude tests; the last two, Spelling and Language Usage, are best described as achievement tests. The speeded test is Clerical Speed and Accuracy; it is administered in two parts, the first of which is unscored and constitutes practice for the second part.

Design of the Computerized Adaptive Edition

The "design" of a computerized adaptive test encompasses several important technical issues: First is the broad issue of the general technical approach to take. Most recent adaptive test development has employed item response theory; the Adaptive DAT is no exception.

To summarize the design features of the Adaptive DAT: the battery includes seven adaptive power tests, intended to be alternate "forms" of the printed edition tests. Each test is individually administered by choosing IRT-calibrated items from a bank consisting of all the items in Form V of the printed edition. All of the items have been calibrated using the Rasch model. Each adaptive test uses Owen's Bayesian sequential updating procedure to estimate the examinee's ability after answering each question. Once the ability estimate is updated, a modified maximum information item selection procedure is used. Each adaptive test terminates when its length is half the number of items in the counterpart test in the printed DAT.

Empirical Research

The Computerized Adaptive Edition is intended to be used interchangably with the printed forms of the Differential Aptitude Tests. For ease of interpretation of test results, this made it desirable that the adaptive test use the same norms as Forms V and W of the printed edition. To justify this, it was necessary to establish a high degree of correspondence between the adaptive and the printed editions, and then to equate the adaptive DAT test scores with the printed test.

Establishing the correspondence of the two different modes of administration meant demonstrating that the two correlated highly, and had similar factorial structures. Equating the two meant deriving transformations that would permit expressing the adaptive test scores as equivalent raw scores of the printed edition.

To accomplish these two purposes, two field tests were conducted, one in the Fall of 1985, and one in the Spring of 1986.

Results

Correlation Analyses For the seven adaptive tests (i.e., every test except Clerical Speed and Accuracy) the correlations across mode ranged from .78 to .88, with a median correlation of .85 . The highest correlations were for the Verbal Reasoning, Numerical Ability, Spelling, and Language Usage tests, for which all correlations were .85 or higher. The lowest adaptive test correlations were those of the three pictorial tests: Mechanical Reasoning, Space Relations, and Abstract Reasoning; their correlations ranged from .78 to .82.

By far the lowest correlation across modes of administration was observed for the Clerical Speed and Accuracy test, where the correlation was .33 . (In a

separate analysis, the reliability of the computerized CSA test was estimated at .85 using the alternate test method. Reported uncorrected estimates of the printed CSA test reliability range from .77 to .93, with a median of .86 (Bennett, Seashore & Wesman, 1982)).

Factor Structures The factor analysis extracted four factors: Verbal Information, Figural Reasoning, Mechanical Reasoning, and Perceptual Speed. Comparisons of the factor structure of the printed edition with that of the Computerized Adaptive Edition indicated that the two batteries were nearly identical. In a separate analysis, to be published elsewhere, researchers reported that the correlation of the printed DAT battery with that of the computerized adaptive one was approximately .97 -- an extraordinarily high degree of similarity given the different modes of test administration.

Summary

The results of the two field tests show a high degree of correspondence between the computerized adaptive DAT tests and their printed Form W counterparts, but little correspondence between the computerized and the printed Clerical Speed and Accuracy (CSA) tests. This discussion will deal first with the seven adaptive power tests, and last with the CSA tests.

The correlations of the seven power tests across modes of administration were high enough to consider the computerized adaptive and the printed editions of the DAT as alternate -- but of course not parallel -- tests. This is supported by the results of the factor analysis, which show a very high degree of similarity of the two batteries, in terms of their patterns of factor loadings. For practical purposes, the patterns of factor loadings of the computerized tests and the printed tests were identical.

Given its high degree of correspondence with the printed edition tests, the Computerized Adaptive Edition could be considered psychometrically equivalent to it.

The two different modes of administering the CSA test, however, did not correlate highly enough to justify equating. Both the correlation analysis and the factor analysis indicate that the computerized CSA test is measuring a somewhat different variable than the printed version. Additional results, not reported here, bear this out. The low correlation between the two CSA tests cannot be attributed to content differences, because the items are identical except for order. The difference probably lies in the different tasks involved in responding to CSA items on the computer screen rather than on an answer sheet. More research is needed into the explanation of these observed differences, and their implications for predicting examinee behavior.

The development and research into the Computerized Adaptive Edition of the DAT will be fully reported in a forthcoming technical report of The Psychological Corporation. That report will be in the form of a supplement to the technical material on the printed editions of the DAT. It will be intended to address the documentation requirements of both the Standards for Educational and Psychological Testing (American Educational Research Association et. al, 1985) and the Guidelines for Computer-Based Tests and Interpretations (American Psychological Association, 1986).

* * * * * * * * * * *

CONCEPTUAL AND PRACTICAL GUIDES IN THE DESIGN OF SIMULATION EXERCISES: IN-BASKETS, ROLE-PLAYS, AND LEADERLESS GROUP DISCUSSIONS

Steve Sonnich

San Bernardino County, California Personnel

Purpose

The purpose of this paper is to provide practical guidelines, based upon theoretical concerns, in the design of content-valid simulation exercises. The intent is to provide a framework from which those who are relatively unfamiliar with the design of these types of simulation exercises, can begin to learn.

Conceptual Framework

The Behavioral Consistency model proposed by Wernimont and Campbell (1968) has served as a theoretical base for work sample, or simulation tests (Schmitt and Ostroff, 1986). This model suggests that tests should be constructed to reflect a point-to-point correspondance between predictor and criterion. The authors suggest that if one is interested in predicting job behavior, then work sample tests should be designed which simulate important aspects of the job. Simulation exercises such as Role-Plays, Leaderless Group Discussions, and In-Baskets are vehicles suited to eliciting observable test behavior that is consistent in content and proportion with job behavior. While this model is certainly rational and useful as a theoretical underpinning for simulation exercises, it should not be interpreted as the conceptual basis for blanket acceptance of content validity as a stand-alone validation strategy.

The following discussion is intended as a brief overview of a long standing discussion in the psychological literature regarding validity. Because of the problems associated with small sample sizes and unreliable criterion measures, an assumption made here is that personnel professionals will often rely on content-validity as a stand-alone validation strategy. The question becomes, under what conditions will content validity alone suffice as the sole validation strategy of a test?

Validity

"The concept refers to the appropriateness, meaningfulness, and usefulness, of specific inferences made from test scores. Test validation is the process of accumulating evidence to support any particular inference." (Standards for Educational and Psychological Testing, 1985, p. 9). The meaning of the particular inference made in most employment selection situations is that the measurement instrument result (test score) must differentiate between those candidates who are more, and less suited to perform a job.

The appropriate means by which to gather evidence about this inference is a function of how one interprets the test score. If the test score is interpreted as a sample of characteristics that candidates currently possess, content or construct validity is necessary. Lawshe states:

> If we wish to infer the extent to which a candidate currently possess (a) a relatively simple proficiency that is a component of the job or (b) knowledge required to perform the job (thus to evaluate a present competence), a content validity analysis is indicated. We use a logical procedure that determines the extent to which the behavior elicited by the test is the same or similar to that required by the job or some portion of the job. Usually the procedure is not a mathematical one, although a quantitative approach is available (Lawshe, 1975).

> And, if we wish to infer the degree to which the candidate currently possesses a trait or other characteristic (usually a psychological construct) critical to job performance (thus, to assess an attribute), a construct validity analysis is indicated (Lawshe, 1985, p. 237).

When test scores are interpreted as signs of future performance (aptitude) in which candidates will subsequently undergo training, criterion-related validity is necessary (Sackett, 1987). Test scores which are interpreted as samples to currently perform must not incorporate that which will be subsequently trained (Guion, 1974, Uniform Guidelines on Employment Testing, 1978, Dreher and Sackett, 1981, APA Principles for the Validation and Use of Selection Procedures, 1987).

Content validity is established through test construction. "Content validity refers to the fidelity with which a measure samples a domain of tasks or ideas; it is the degree to which scores on the sample may be used to infer performance on the whole." (Guion, 1974, p. 289). The author suggests that appropriate application of content and construct validity can be viewed as a function of how directly the "job content domain" is sampled. The greater the "inferential leap" necessary to relate test content to job content, the less appropriate content validity becomes. Further, Guion suggests that the "inferential leap" can be viewed along a continuum. At the low end of the continuum are types of tests such as probationary periods and job simulations which would more directly sample the job content domain. At the high end of the inferential continuum, tests assessing general and basic traits would be less likely to directly sample the job content domain.

If a content validity strategy is pursued, great attention must be paid to how one defines job dimensions so that they may be represented proportionally on the test. In addition, one must establish how these definitions are linked to observable job behaviors. In an article on the difference between content, construct, and criterion-related validity (Tenopyr 1977), sounds a cautionary note with resepect to the proper use of content validity.

> If you want to use inferences about test construction to justify inferences about test scores, stay with simple, well defined constructs with easily observable manifestations. (p. 49)

The extent to which a content strategy may be used as the sole basis for test score validity can be viewed as a function of the level of specificity with which the knowledges, skills, and abilities of the content domain are defined. In noting the inconsistencies with which the terms knowledge, skill and ability are defined the APA Guidelines state: "Researchers have frequently called the knowledge or skill related to a small group of tasks an ability. When the ability is defined in this very specific way, content-oriented strategies may be sufficient. When referring to more general abilities such as reasoning or spatial ability, a construct-oriented strategy is likely to be necessary" (p. 19).

By increasing the fidelity with which the abilities are operationally defined, one can decrease the level of inference concerning behavior and the ability it represents. Ultimately, where the line between content and construct validity should be drawn will rest upon case law.

This issue does not appear to have been fully decided by the courts. Dreher and Sackett (1981) suggest that the inferences that the courts will draw from the Guidelines regarding the appropriateness of content validity in various settings are likely to vary from case to case.

In summary, there does not appear to be an overwhelming mandate for content validity as a stand-alone validation strategy (Sackett, 1987), particularly for jobs that are relatively complex. Nevertheless, personnel assessment specialists must design tests for relatively complex jobs which are based upon a content-oriented validation strategy. And, simulation exercises are often more desirable than traditional forms of testing, because they can capture the complexity of the job (Kaman and Benson, 1988). Unfortunately, there are few published "how to" manuals for the design of simulation exercises. What is the practitioner to do? The remainder of this paper is devoted to developing guidelines based upon the conceptual framework and practical techniques that can be followed in the design, administration, and scoring of Role-Play, Leaderless Group Discussions, and In-Basket exercises. These guidelines will be presented separately, but should be thought of as series of mutually dependent steps in constructing simulation exercises.

Guideline 1 - Conduct a thorough task-based job analysis that categorizes behaviors into knowledges, skills, and abilities which form the basis of operational definitions of job dimensions.

Guideline 2 - Determine if you have the resources to successfully complete the project.

Guideline 3 - The test format, content, and administration must allow candidates the opportunity to manifest the targeted dimension behaviors in a manner as close to the job context as possible.

Selecting the Test Type

One must decide which type of test is best suited to assess the dimensions defined in the job analysis, for a specified job content domain. The driving force behind the decision about which type of selection exercise to use is the job analysis. Otherwise, a content validity strategy makes no sense. Selecting the right type of exercise, however, in no way assures that the job dimensions will be assessed. They are formats in which certain types of behavior may be observed better than others, but the content of the exercise and the manner in which the design allows behavior to be manifested is the key to demonstrating the degree of content validity.

Variability in Simulation Exercises

Simulation exercises allow candidates to demonstrate rather than indicate behavior. In multiple-choice style tests candidates are presented with a question and usually four courses of action. The choice indicates how the candidate says he/she will act, but the candidate does not actually manifest the behavior. Variability results when candidates make different choices over many questions.

Simulation exercises differ in that candidates often must put facts together to formulate the question, decide how to act, and manifest behavior. The choice about how to behave is up to the candidate. While the range of behavior is finite, candidates have a high degree of response freedom. With a high degree of response freedom one would expect a high degree of variability.

The stimulus (simulation exercise) should have uniform meaning so that the variability of responses is primarily attributable to candidates and not to the manner in which the information is presented. The information should be structured so that reasonable but inappropriate conclusions can be drawn. Candidates who draw inappropriate conclusions will demonstrate inappropriate behaviors. Candidates who draw appropriate conclusions will demonstrate appropriate behaviors. While these statements are generalities, they indicate how variability can be conceptualized in designing simulation exercises.

Given this conceptualization, the design issue is how to create an exercise with a high degree of response freedom for variability among candidates and yet provide enough structure for reliable assessment. This can be accomplished by considering, during design, how candidates may construe the facts presented, what conclusions they may draw, and how they might behave. By considering the behaviors that might occur, flaws in design can be uncovered so that the exercise more closely approximates the job.

Guideline 4 - Raters must be thoroughly trained in observing and coding behavior into ratings.

The assumption that by operationalizing all aspects of the test development process one can reasonably infer that the test score is valid is squarely contingent upon the reliability of ratings. Without interrater agreement, the rationale for content validity holds no weight. In fact, this issue is so fundamental that Ebel (1979 p. 303) suggests that "content validity" should be called "content reliability." Ratings, subjective judgements based upon job standards, must correlate for one to begin to argue that the targeted measures were accurately assessed. As a result, rater training cannot be overemphasized.

Guideline 5 - The scoring system must be designed to accurately identify high and low performers in terms of job behavior.

In most selection settings, particularly the public sector, candidates are placed in rank order based upon test score. The Guidelines indicate that rank ordering based upon a content valid test should be used only if it can be shown that a "higher score . . . is likely to result in better job performance." Without the empirical relationship that test performance is correlated with job performance (criterion-related evidence), the courts are likely to pay close attention to the fidelity of the scoring system when candidates are rank ordered (Guardians v. Civil Service Commission of New York, 1980). As a consequence, the scoring system must provide standards for raters to apply concerning what is positive and negative dimension behavior. In addition, the rationale for how dimension scores are to be combined should be based upon the job analysis.

The standards for rating test behaviors should be based upon how those behaviors would result in positive or negative outcomes on the job. Positive and negative outcomes can be gathered through critical incident data and subject matter expert concensus. Rating scales can then be created which allow raters to assign scores to behaviors they have classified in terms of job dimensions.

* * * * * * * * * *

DEVELOPMENT OF JOB-RELATED MEDICAL STANDARDS/GUIDELINES FOR SELECTION OF APPLICANTS AND EVALUATION OF INCUMBENT PERSONNEL

Deborah L. Gebhardt & Carolyn E. Crump

Advanced Research Resources Organization
A Group of University Research Corporation
Chevy Chase, Maryland

Job-related medical standards and guidelines promote safe and effective personnel placement and lower accident/injury rates. They provide for the acceptance of qualified handicapped applicants for specific jobs and aid in the development of reasonable accommodations for these applicants. Medical standards and guidelines are concerned with the degree of impairment within a body system and whether a specific level of impairment limits an individual's capacity to perform critical job tasks.

A medical examination should be an evaluation of an individual's ability to perform job tasks effectively and safely. In order to ensure that the examination takes into account both the job tasks and environmental working conditions, the examining physician should be provided with guidelines that aid in assessing the health status of an individual in relation to the requirements of the job. The most useful guidelines are those which outline the levels of severity of the medical diseases and conditions that affect performance of the critical job tasks.

The approach to determine medical standards and guidelines and the database described in this paper have been developed by Advanced Research Resources Organization (ARRO) through a programmatic research effort that has spanned an eight-year period. Research by Gebhardt and Crump (1982, 1983, 1984, 1986, 1987a, 1987b) has resulted in a methodology that uses task specific information to provide accurate and comprehensive medical standards and guidelines. The methodology designed by ARRO provides backup data showing the relationship of each critical job task to a specific disease. Use of this methodology results in a product that is based on job-related criteria, that is legally defensible, and that is targeted to the physician. The medical standards and guidelines developed for the auditory, cardiovascular, endocrine, gastrointestinal, genitourinary, integumentary, musculoskeletal, nervous, respiratory, and visual systems are formatted into a Physician's Manual.

Methodology

ARRO's unique methodology links the job requirements obtained in a job analysis to the medical standards/guidelines. It can be used to determine both selection and retention medical standards and guidelines. This methodology involves a systematic approach of identifying the severity of a specific disease/condition that limits or precludes safe and effective performance of critical job tasks. Medical specialists (e.g., cardiologists, orthopedists, neurologists, occupational physicians) and ARRO staff use a three-stage approach to identify the standards.

First, a job analysis is completed that identifies the critical job tasks and clusters them into specific categories (e.g., push, climb, comprehension, vision). The environmental working conditions in which the critical job tasks are performed, are also identified during the job analysis phase. The ergonomic parameters (e.g., heights, weights, lighting) of the work setting are determined through on-site visits and data collection.

Second, data concerning the accidents/injuries and compensation costs are analyzed and compared with the ergonomic and environmental data related to the critical job tasks. These data are used to provide information about the tasks which have accounted for the greatest number of accidents/injuries and/or compensation costs. Further, the nature and severity of the injury or illness, body part injured or affected, location of accident, and probable cause are analyzed in relation to the tasks being performed (Gebhardt, Cooper, Jennings, Crump, & Sample, 1983; Gebhardt, Crump, & Frost, 1987).

Third, the job analysis, accident/injury, environmental, and ergonomic data are consolidated and matched to the same type of information contained in ARRO's computerized Medical Database. For new jobs this information is submitted to the ARRO medical model in which medical specialists use a rating system to determine the level of severity of a disease or impairment that will impact job performance. The rating system utilizes scales, developed by ARRO and medical specialists, that define diseases/conditions in terms of symptoms, function, and medication. These rating scales and the rating procedure provide the basis for evaluating the severity of the diseases/conditions that impact performance of the critical job tasks (Gebhardt et al., 1983; 1986; 1987).

Medical scales have been developed for the diseases in each body system (e.g., cardiovascular). Presently, ARRO has developed individual medical scales for over 250 diseases across the ten body systems. Each disease scale is defined by levels of severity which are described in terms of the symptoms, medication, and function associated with a specific level of severity. These scales are continually updated to reflect current medical advances.

Separate meetings for each medical specialty (e.g., orthopedics) are held. Each panel of specialists is given the consolidated job analysis and accident/injury information, along with a briefing about the job under discussion. For jobs previously analyzed by ARRO or jobs in which the critical tasks can be matched to similar critical tasks in other job titles in ARRO's Task Bank, the medical standards for a task are initially generated from ARRO's computerized Medical Database. These are reviewed by the medical specialists to ensure that specific job conditions (e.g., environmental, frequency) have not been overlooked which have an effect upon a particular disease/disorder. For new jobs or new critical tasks within a previously studied job, the physicians rate each critical job task on each disease. This is followed by a discussion of each task within each disease/condition to arrive at a consensus of the level of severity that precludes effective task performance.

At the completion of the medical meetings, the level of severity that precludes safe task performance for each disease and condition will have been determined for each critical task. This information is input into the Medical Database and provides the rationale for determining the final selection medical standards/guidelines. During this process, the determination of the standards/guidelines for evaluating incumbents is also undertaken. The determination of the retention standards takes into account job rank (e.g., sergeant, captain) and the progression of a disease/disorder.

Physician's Manual

Following the identification of the level of disease severity that precludes safe job performance, a Physician's Manual is developed (Gebhardt, 1983b). This Manual provides the examining physician with background information related to the job duties and an itemization of the level of severity of the diseases and impairments that would disqualify an individual from the job. The Manual includes (1) a description of the job as determined from the job analysis; (2) the disqualifying level of severity for each disease/condition in each body system, as well as the acceptance level of a disease/condition; and (3) an indication of areas that necessitate additional evaluation by a medical specialist (e.g., cardiologist).

Two Physician's Manuals can be developed, one for selection and one for evaluation of incumbent personnel. The first Physician's Manual is used for screening applicants for an entry-level position. The second Manual is used to evaluate incumbents and may be targeted to a variety of positions within a job classification.

Light Duty Assignment

The ARRO Medical Database can be used to establish a system that identifies the tasks an individual can perform after returning from an injury or illness (Gebhardt & Crump, 1984). This system can help the employer assign an individual to specific job tasks in their present job and parallel tasks in other job titles for which they are qualified. The use of such a system provides the employer with a method to identify the percentage of critical tasks within the job that the injured/ill employee can perform. The employer can therefore determine whether the number of tasks an employee can safely perform is adequate to warrant returns to the job.

Application of Methodology to a Variety of Jobs

Once the level of severity of a disease/condition that precludes safe task/job performance has been identified, this information can be transported to other similar jobs. The transportability of the medical information is based on a similarity analysis that incorporates identification of critical job tasks, environmental conditions, and ergonomic parameters. This information is then matched with previously analyzed jobs in the ARRO Task Bank to establish job and task similarity. Following this matching procedure, the medical standards/guidelines per task and per job are generated from the Medical Database and formatted into the Physician's Manual. These procedures comply with the Federal Uniform Guidelines for selection and take into account other statutes such as the Rehabilitation Act of 1973 and Age Discrimination in Employment Act.

AUTHOR INDEX

223