8th Annual

Military Testing Association Conference

PROCEEDINGS

Hosted by
NAVAL EXAMINING CENTER

Waukegan Inn
Waukegan, Illinois
19-22 September 1966

1966 MTA COMMITTEES

President

Captain R. M. Hayes

Chairmen

Cdr R.J. Gray
Mr. C. S. Winiewicz
Mr. C. J. Macaluso

Coordinator

Mr. J. W. Crediford

Steering Committee

Mr. C. S. Winiewicz, Chairman
Cdr R. J. Gray
Mr. C. J. Macaluso

Public Relations/Protocol

LCdr L. Forte (W), Chairman
Mr. J. C. Houtz
R. W. Thurlow, JOC
LCdr J. Hennessey

Program

Lt D. Y. Acosta (W), Chairman
Mr. C. S. Winiewicz
Mr. C. J. Macaluso
Mr. L. W. Johnston
Mr. W. R. Birdsall
Mr. J. R. Rush
Mr. E. A. Dukerschein

Physical Arrangements

LCdr J. D. Lee, Chairman
Mr. C. A. Hudson
LtJg C. P. Byrnes
Lt M. L. Keating (W)

Transportation

LtJg D. E. Fadden, Chairman
Ens M. R. Holder

Registration

Lt C. E. Gengenbach, Chairman
Mr. R. L. Guy
J. E. Cook, YNCS
LtJg V. W. Decain

Hospitality

Ens J. L. Boyd (W), Chairman
LtJg R. L. Allison
Mr. P. J. Buranowski

Entertainment

Lt S. V. Hart, Chairman
LtJg R. L. Haughey

Dinner/Banquet

Lt J. I. Marrone, Chairman
Mr. H. R. Gagnon
Mr. V. H. Beggs

FOREWORD

The Eighth Annual Conference of the Military Testing Association was held at the Waukegan Inn, Waukegan, Illinois, 19-22 September 1966. The Naval Examining Center was the host organization.

All papers received have been included in these proceedings. The speakers who spoke extemporaneously, or who did not forward their papers, have not been forgotten and will not be forgotten by the conferees, for their addresses, also, were memorably superb. The conference was characterized by the great generosity of the speakers and the cooperative attitude of the members. The theme of the conference was "On-going Human Appraisal." If we apply the words of the theme to the speakers themselves, we honestly state that their work must be appraised as superlative.

The standards of the conference were set in the opening remarks. Captain V. J. Soballe, USN, Commanding Officer, U. S. Naval Training Center, Great Lakes welcomed the group. The invocation was provided by Captain R. F. McComas, CHC, USN. Captain R. M. Hayes, USN, Commanding Officer, Naval Examining Center welcomed the group. Colonel James C. Donaghey, USA, Commanding Officer, Enlisted Evaluation Center, Fort Benjamin Harrison spoke of the work accomplished there during the year and of their future accomplishments. Colonel J. H. Ritter, USAF, Commanding Officer, Personnel Research Laboratory, Lackland AFB, Texas contributed an inspirational message for the group. Captain G. T. Murati, USCG Commanding Officer, USCG Training Center, Groton, Conn., using an analogy based on dead reckoning, estimated the course of future testing in the U.S. Coast Guard.

Dr. Harry Greer, jr, Captain, USN Ret., former commanding officer of the Naval Examining Center and founder of the Military Testing Association, was introduced by his former shipmate, Captain R. M. Hayes, USN. His keynote address truly set the policy for an interesting conference.

To continue acknowledgment in this foreword would mean listing individually every name included in these proceedings. If space permitted, it would be desirable to do so, but a grateful "Thank You" is extended each of these hard-working individuals.

It is understood that the opinions freely expressed in these papers are those of the authors. They are not to be construed as official or in anyway representative of the U.S. Armed Services.

MILITARY TESTING ASSOCIATION CONFERENCE
THEME: ON-GOING HUMAN APPRAISAL

CONFERENCE CHAIRMEN:	COMMANDER R. J. GRAY, C. S. WINIEWICZ, C. J. MACALUSO
CONFERENCE COORDINATOR:	JOHN W. CREDIFORD
PRESIDENT MTA 1966:	CAPTAIN R. M. HAYES, U. S. NAVY
HOST ORGANIZATION:	NAVAL EXAMINING CENTER
PLACE:	WAUKEGAN INN, WAUKEGAN, ILLINOIS
TIME:	19 - 22 SEPTEMBER 1966

MONDAY - 19 SEPTEMBER 1966

0800-1700	Army Meetings
1300-1700	Registration - Waukegan Inn
1700-	Cocktail Party and Buffet at the Great Lakes Commissioned Officer's Club

TUESDAY - 20 SEPTEMBER 1966

0900-0905	Invocation - CAPTAIN ROBERT F. McCOMAS, CHC, U. S. Navy
0905-0915	Welcoming Address - CAPTAIN R. M. HAYES, Commanding Officer, Naval Examining Center
0915-1000	Introduction of Special Guests - C. S. WINIEWICZ, J. W. CREDIFORD, C. J. MACALUSO, Naval Examining Center
1000-1030	COFFEE BREAK - OPEN DISCUSSION
1030-1200	Keynote Address - State of the Art - Evaluation in Planning: A Recommended Process for Military Planners - CAPTAIN HARRY H. GREER, U. S. Navy (ret)
1200-1315	LUNCH
1315-1345	NEC Outlook - CASIMER S. WINIEWICZ, Naval Examining Center
1345-1405	Commentary - Applications of the Computer in Enlisted Classification - EDMUND THOMAS, U. S. Naval Personnel Research Activity, San Diego, California

TUESDAY - 20 SEPTEMBER 1966 (continued)

1405-1425	Commentary - Assessing Human Capabilities through Equipment Oriented Tests - DR. LEONA EGAN, Personnel Program Support Activity, Washington, D. C.
1425-1445	Ultimate Use of Computers in Test Interpretation - GLEN ANDERSON, National Computer Systems, Minneapolis, Minnesota
1445-1515	Computer Diagnostic Profiles Derived from the MMPI - DR. JOHN PEARSON, Mayo Clinic, Rochester, Minnesota
1515-1530	COFFEE BREAK
1530-1715	Presentation of Papers - Chairman: ROBERT L. GUY, Naval Examining Center
1535	A Questionnaire to Improve a Testing Program - J. E. PARTINGTON, U. S. Army
1555	The Position of Army MOS Proficiency Testing in an Operation's Research Context - DR. RAYMOND O. WALDKOETTER, U. S. Army
1615	Application and Utilization of Cutoff Scoring Procedures - DALE R. BAKER, U. S. Army
1635	Army Experience with Field Scoring of Performance Tests - CHARLES D. ROBERTS, U. S. Army
1655	An Approach to the Measurement of Supervisory Skills - WILLIAM W. WANCE, U. S. Army
1715-	FREE EVENING

WEDNESDAY - 21 SEPTEMBER 1966

0900-1000	Abstract of Presentation - DR. CHARLES A. POUNIAN, Personnel Officer, Chicago Civil Service Commission
1000-1030	COFFEE BREAK - OPEN DISCUSSION

WEDNESDAY - 21 SEPTEMBER 1966 (continued)

1030-1200	The DDC Mission and Services - LT COLONEL VOYA N. SKAKICH, DCS, Cameron Station, Alexandria, Virginia
1200-1315	LUNCH
1315-1430	Systems Analysis Applied to Personnel Research - DR. J. M. KEENAN, Stanford Research Institute, Menlo Park, California
1430-1445	COFFEE BREAK
1445-1700	Presentation of Papers - Chairman: VERNON H. BEGGS, Naval Examining Center
1450	A Rapid Solution for Multiple R and Beta Weights with Checks - JOHN S. BRAND, U. S. Army
1510	A Note Concerning Correlations with Weighted Composites - JOHN S. BRAND, U. S. Army
1530	Cross Validation Study of MOS 23G40 - (Hercules Fire Control Mechanic) PAUL P. FOLEY, U. S. Army
1550	Measuring Troubleshooting Skills of Maintenance Personnel by Paper and Pencil Tests - JOHN L. FINUCANE, U. S. Army
1610	Current Trends in Bandsman Performance Tests - ANTHONY STRIGARI, U. S. Army
1630	Do Schematics Help? - FRED B. HONN, U. S. Army
1650	Intra-individual Variance as an Appraisal Technique - CLAUDE BRIDGES, U. S. Army
1700-1900	FREE TIME
1900-2000	Banquet - Toastmaster: HERMAS R. GAGNON, Naval Examining Center

WEDNESDAY - 21 SEPTEMBER 1966 (continued)

2000-	Banquet Speaker - "Tribulations of Testing" - DR. PHILIP ASH, Research Assistant to Vice President, Inland Steel Company, Chicago, Illinois

THURSDAY - 22 SEPTEMBER 1966

0900-1000	Psychological Assessment of Executive Personnel - DR. IVO MERSMANN, United Airlines
1000-1030	COFFEE BREAK - OPEN DISCUSSION
1030-1200	Human Engineering Methods of Appraisal - DR. ROBERT L. MALONE, Fry Consultants, Inc., Chicago, Illinois
1200-1210	Steering Committee Report
1210-	Closing Remarks

KEYNOTE ADDRESS: EVALUATION IN PLANNING: A RECOMMENDED PROCESS FOR MILITARY PLANNERS

Captain Harry H. Greer, jr. USN (RET.)

Dr. Greer briefly reviewed the "state of the art" as he found it upon reporting to the U.S. Naval Examining Center in December 1958. He gave as one reason for founding the MTA as a mechanism for material benefit and profit through regular professional exchange. He went on to note the recent strides in military testing particularly since the introduction of the digital computer to scoring and analysis. It was pointed out that the future may see more emphasis on interest and personality instruments although there has been some criticism of these forms of tests. The impact of new technology was surveyed noting, among other things, the ability of a computer to randomize item selection and the use of terminals in test administration. In conclusion Dr. Greer discussed the future of testing.

BLANK PAGE

NEC BACKGROUND AND OUTLOOK

Casimer S. Winiewicz
Naval Examining Center, Great Lakes, Illinois

INTRODUCTION:

During the February 1959 world-wide Navy Advancement Examination administration period, 778 subjects, representing almost the entire spectrum of GCT (general intelligence) ability and three different levels of achievement, were first introduced to the porta-punch card as a method of taking objective type examinations.[1] Since then, additional backup studies composed of approximately 3,000 subjects and different configurations of EDPM and EAM equipment finally culminated in the abandonment of the traditional testing approach and the punched hole card was introduced to the entire U. S. Navy. In August 1959 almost a quarter-million candidates qualified for advancement took their competitive examinations on the porta-punch answer card. Since the initial introduction of this unique answer card, several million examinees have taken their examinations in this form to date.

The substitution of the card for the traditional marked sense answer sheet required no additional time on the part of the examinee and created perhaps the largest Source Data Automation program in the world. Every Navy examinee is now a highly motivated "Key Punch Operator", because of additional money and advancement being dependent upon his responses, and because he creates direct input to the Examining Center's Data Processing System - wherever he is located throughout the world.

BACKGROUND

During a typical year approximately 1700 different examinations are developed to reflect the scores of technical skills and fields of knowledge required in today's Navy. They are administered annually in February and August under completely controlled conditions. This entails shipping and accounting for three-quarters of a million examinations, as

The opinions and assertions expressed in this paper are those of the author and are not to be construed as official or reflecting the views of the Navy Department or Naval Service at Large.

[1]Results of which were presented by C. S. Winiewicz and C. J. Macaluso at the American Psychological Association Convention on September 7, 1959 at Cincinnati, Ohio.

well as processing their results and promulgating the information back to the examinees.

In order for a candidate to advance in rate (occupation), he must be qualified in all respects, recommended by his Commanding Officer and successfully compete on a service-wide advancement examination. Each examination is weighted to determine the number of items which are to be included for each subject matter area breakdown under the applicable job level for that rate. Each instrument is composed of two major parts: professional which pertains to the specific job area (usually 120 - 4 option multiple-choice items) and the military (usually 30 - 4 option multiple-choice items). Because of the lack of rigidly standardized examinations (as contrasted with commercial instruments) and the fact that each examination is utilized only once on a given population to preclude compromise or collusion, and also to maintain the flexibility of the instrument to reflect current technological changes, some control factor is required to tie the test into previous test results to insure comparable standards. Pre-testing is not feasible because of the magnitude of the overall program and the relatively tight scheduling required in the continual development and processing of all examinations.

The control factor that is utilized to tie in present examinations with previous results is to control the design of each new examination by including 50% of the items previously used with known statistical characteristics obtained through item analysis, and a constant check of separate periodic research studies on the examinee population. The remaining 50% of the items are newly developed for use in the current examination. Consequently through a rigid control of item and statistical analyses and related special studies, the test instrument assumes greater standardization and control than some of our so called standardized commercial instruments on the market today.

In determining an equitable cut-off score for each of the 1700 different examination populations that will dichotomize our groups into pass - fail categories and still maintain comparability between different examinations and examining periods, a multitude of factors must be taken into consideration, such as needs of the service at large, budgetary problems, and above all the qualifications and performance of each candidate. The raw scores are converted to standard scores (via linear and non-linear transmutations) with a range from 20 - 80, with a mean of 50 and a sigma of 10. This scale coincides with t scores and takes advantage of all the properties of the normal curve.

Final advancement, assuming the candidates have fulfilled all the necessary prerequisite requirements and have

attained a passing score on the examination is determined by one of two methods depending upon the criticality of the occupation under consideration. First if the candidate happens to compete in an occupation whose skills are in demand (more vacancies available than qualified personnel to fill them), the examination passing score qualifies the candidate for advancement. However, if the candidate competes in an occupation where the number of vacancies are small (more qualified personnel available than existing vacancies), final selection of those authorized to be advanced is made on the basis of relative standing on a final composite score consisting of the following five factors, along with their maximum values: (1) Performance Factor (50), (2) Length of Service (20), (3) Time in Rate (20), (4) Number of Awards (10), and (5) Examination Grade (80), giving a maximum composite score of 180.

This system provides an equitable opportunity to compete for advancement, under completely controlled test conditions for the number of authorized advancements regardless of the place that an individual may be stationed throughout the world, his present duty assignment, or the vacancies or surpluses that exist in the local commands. The Navy maintains approximately 3200 activities (ships or stations) located in all parts of the world, and classifies men into over 65 different occupational skills divided into 6 different levels of ability.

The measure of any testing system, however, rests on the reliability and validity of the instruments employed and in this case how well they will identify the individuals in respect to performing their assigned duties in a satisfactory manner. For purposes of statistical analysis this criterion is expressed as some kind of score such as school grade, proficiency on job through rating or some other measurable indicant of effectiveness. The importance of a criterion is expressed by the following quotation from a Secretary of Defense report.[2]

"The fundamental basis of all selection and classification on methods and procedures is the establishment of adequate criteria. Without such criteria the selection and classification of personnel, however refined the methods of evaluation become, are meaningless. The establishment of adequate criteria presumes, first, that the personnel requirements are known. Often this is stated in terms of numbers required for a particular duty or job; i.e., radar maintenance, engine mechanics, pilots, tank operators, etc. These require-

[2]Report of Working Group on Human Behavior Under Conditions of Military Service, Office of the Secretary of Defense, HBM 200/1. 1951.

ments are almost never stated as they should be in terms of the kinds of traits, capacities, skills, and personalities essential to the job."

Validity coefficients for Navy advancement examinations against the criterion of on-the-job performance measures, obtained by supervisory ratings, fall between the average to high range for coefficients of this type. Sixty per cent of these validities range from r = .20 to r = .70; values which attest to the validity of these examinations when one considers the nature of the criterion and the fact that testing personnel have experienced that correlations between test scores and ultimate external criteria rarely extend beyond .40, and are generally concentrated in the .20 to .30 range. (See Validity Exchange Information, Personnel Psychology).

It is the function of the U. S. Naval Examining Center within this framework to; construct examinations, ship and receive them back for processing, account for every examination used or unused, evaluate statistically all examinations, maintain the integrity of the examining system, select the most qualified candidates for advancement, and finally to continually strive to improve the entire system.

EFFECTS OF CHANGE TO DATE

1. Complete flexibility and freedom for expansion. The original change to the punch card was associated with a medium size digital electronic computer (IBM 650). The 650 was replaced by an IBM 1401 solid state card computer which in turn has given way to an UNIVAC 418 tape system, which shortly will be interfaced with an NCS optical scanner.

2. The advancement examination for pay-grades E-4 and E-5 have been based upon professional requirements alone, so that broader coverage of the basic qualifications are covered. The military portion has been made into a separate 100 item examination that will be administered on a quarterly basis beginning in January, 1967.

3. Previous manual operations such as ordering of examinations, shipping procedures and accountability are now part of the computer system.

4. Greater control has materialized through a more uniform and standardized system, enabling the center to become an integral part of the overall Navy Manpower System.

5. Additional programs such as: administration of college board examinations for the NESEP program, officer selection (LDO-INT), Warrant and Reserve Warrant, Defense

Officer Retention Examination, Aptitude and Foreign Language Proficiency Examinations, and other special officer and school selection programs have been incorporated into the system.

6. Profile analysis and feed back information to examinees on their strong and weak points on the different areas of the examination and the final multiple advancement information have been provided the candidates.

7. Additional research studies on the examinee population, service schools, the test instrument, and more basic exploratory pure theoretical approaches.

8. Greater emphasis on a more sophisticated statistical approach to evaluating the examinations and their analysis of items.[3] No longer are short cut estimates based on other estimates obtained from tables the most expedient method to analyze test instruments and their items.

9. Minor changes may be quickly incorporated into the system without disruption of present operations.

ANTICIPATED FUTURE CHANGES

1. Like all other services NEC will have a greater percentage of civilian personnel due to the civilian conversion program.

2. NEC will become an integral part of Project Compass for the Great Lakes area by the computer assistance provided in the classification area.

3. Some special school and officer selection programs will be converted from punch cards to answer sheets so that the capability of the optical scanning device may be employed. After a series of experimental and feasibility studies are completed, thought will then be centered upon placing the entire naval advancement program on the optical scanner.

4. The elimination of manual categorizing of item parameters by Educational Specialists and effects of up or down grading will become a function of automatic computer output.

5. Elimination of item bank files by placing the items plus their related statistical and occupational codes for

[3] Winiewicz, C. S., NEC RESEARCH NOTE 60-1 Statistical and Item Analysis Procedures.

selection on "magnacards" for items and "magnavue" cards for illustrations. They would be placed in machine storage and selected out according to a programmed key based on the needs of the examination under construction.

6. Complete control of the entire system by the continual updating of a master tape containing all pertinent information on all candidates in the entire U. S. Navy. This procedure will eliminate the preparation of various documents that were previously necessary and will make long and short range prediction on personnel matters with a great degree of accuracy.

7. Finally, card handling and the intervention of human operators will be held to a minimum with a corresponding decrease in possible errors.

CONCLUSION

The present system has met with wide acceptance from all candidates who have participated in taking the various navy examinations. This does not imply that ours is the only approach, but that it is a system pioneered by the Navy and found to be more successful than we originally hoped.

With the advent of optical scanning and document reading devices, it is possible that some aspects of the punch-card approach will become obsolescent, however NEC by that time should have enough empirical and experimental evidence to make a smooth transition into these new frontiers.

In conclusion, with the rapid technological changes in both EAM and EDPM equipment, the speed with which data can be treated and analyzed is only restricted by the ability of man to imagine and hypothesize.

APPLICATIONS OF THE COMPUTER IN ENLISTED CLASSIFICATION

Edmund D. Thomas
U. S. Naval Personnel Research Activity,
San Diego, California

Our research in the area of personnel selection and classification is becoming increasingly important as more technically advanced equipment continues to be introduced in the modern Navy. Unlike the early sailor, who might be called a "jack-of-all-trades," today's enlisted man is often a specialist who is responsible for one or more types of complex equipment. Thus, accurate determination of the "right" job for each man is important, not only to maximize his value to the Navy, but to increase his feeling of success within the assignment--a real problem for those involved in selection research.

Actually, selection in itself is not a particularly difficult problem in the Navy. Using our present tests, consisting of the five Basic Battery Tests and four special tests, and having an available pool of over 120,000 recruits each year, it's a fairly straightforward task to determine which men can best profit from additional vocational training. A much more complex and real problem is that of assigning a man with sufficient scholastic aptitude to one of more than 70 possible Navy basic schools in such a way that his chances for success and ultimate usefulness to the Navy can be realized. Our efforts to improve classification can not only provide more and higher caliber basic school graduates, but also can reduce the number of school failures with attendant costs in training time and facilities and in lowered morale.

Until a few years ago our research in the area of personnel classification was oriented primarily toward the development, evaluation, and revision of selection and classification tests. Among our recent accomplishments are the construction of Form 7 of the Basic Test Battery, referred to as the BTB; continued monitoring of the effectiveness of BTB within each of the Navy's basic schools; developing and validating a vocational interest type test; and conducting various methodological studies in classification testing and test construction.

In general, the approach taken in this research has been toward providing measures of different aptitudes which could permit us to better classify men into the various

Navy jobs--ratings. At present a recruit may take from five to nine different classification tests designed to measure quite different aptitudes or achievements.

With the increasing availability of digital computers (which Dr. Greer mentioned this morning) the direction of our research is changing and we are becoming involved with devising and applying techniques which can more effectively make use of this test data in classification. A particularly challenging area of investigation has been in the development of Project COMPASS--which is sort of an acronym for computer assisted assignment. COMPASS is a product of several years of effort by the Personnel Measurements and Statistical Departments of the Personnel Research Activity in San Diego, and was accomplished with the cooperation of the Classification Staff Personnel at the Naval Training Center in San Diego. Its development has included over a year of computer programming, debugging, trial runs, and trouble shooting. COMPASS has now been operational for about ten months at the Naval Training Center in San Diego and upwards of 50,000 men have been processed using the COMPASS Programs.

Basically, COMPASS has been designed to modernize or "streamline" the classification process. It relieves the classification interviewer--who works with the recruit toward recommending suitable Naval assignments and jobs--of many routine clerical tasks and allows him more time to concentrate on qualifications and expressed motivations in line with the needs of the Navy. COMPASS also assists the interviewer by indicating which basic schools or types of on-the-job assignments in the Fleet that each recruit is qualified for.

There are a number of reasons why the Navy became interested in the possibility of developing a computerized approach to classification. One obvious reason was its potential for <u>filling school quotas</u>. During certain times when the quality of men entering the Navy is lower than desired, school quotas may not be met or else requirements are waived for men of marginal aptitude. This results in less efficient use of training facilities, higher dropout rates, and possibly an insufficient number of trained men in some ratings. Another consideration was that not only would we be able to fill assignment quotas, but at the same time we could <u>improve the overall quality</u> of men assigned to schools. Through the use of a computer all test information can be combined in more mathematically complex ways than we are presently using, resulting in predictions which are more validly related to school achievement.

Also, there is the immense task required in classification. The problem of assigning in an optimal way the maximum number of school qualified men is a very complex one. It is, in fact, impossible for an individual classifier to give the best mathematical weights to each of the various selection factors. Through the use of the computer, the necessary computations can be made at a remarkably rapid rate to obtain the overall best combination of men and jobs.

Although time is limited I would like to briefly outline the four operational programs comprising COMPASS, beginning with Program PROFILE, the first in the chain of programs. PROFILE provides a listing of those schools for which the recruit is eligible on the basis of aptitude test scores and school guarantees made as a condition of enlistment. Each recruit's Profile Card lists the schools for which he is eligible along with sums of selector scores used for these schools. This listing is used by the classification interview as a guide in discussing school assignments with the recruit. PROFILE thus assures the accurate addition of test scores required for school selection and also may help the undecided recruit by suggesting school possibilities. Following the interview the recommendations of the interviewer are then punch-card coded for the second phase of COMPASS--Program SCREEN.

SCREEN checks for possible errors in coding, interviewer recommendations, keypunching, etc., for each recruit. In a sense, it insures that each man is qualified for the schools for which he has been recommended.

An unexpected outcome of COMPASS came during the development of the SCREEN Program. Each of the 70 or more Navy schools has its own set of requirements, such as minimum test scores, color vision, security clearance, and special physical qualifications, for example. It is not surprising that we found that a small, but significant, number of men were being assigned to schools for which they were not qualified. As a result of this discovery, SCREEN was adopted by the Naval Training Center in San Diego before the other COMPASS Programs were completed, because of its value in spotting these errors.

Our first operational assignment program, called OPERATE, maximized the sum of scores for all available men, for all schools, within the limitations imposed by quotas in special enlistment programs. OPERATE filled school vacancies using a fairly complex mathematical model. In effect, it considered every man for every school and made assignments to these schools on the basis of summed aptitude test scores.

As you may imagine, this program proved somewhat complex and costly in operation, and a new approach to school assignment was sought--to reduce costs, and also to get more men into their first choice of duty. This lead to the development of Program SKOOL which is currently operational as our assignment program.

With Program SKOOL, which is a type of sorting and matching program, men who enlisted with a guaranteed assignment to a service school are assigned before those without a guarantee. The several Navy school guarantee programs have been assigned positions based upon relative priority and importance. Within each group the men are assigned to their first recommendation (choice) and if the quota is exceeded, the men with the lower selection test scores are pulled out and are considered for their second choice. If the available quotas are again filled by better qualified men, the man's third, fourth, and fifth choices are considered in turn. Non-school-guaranteed men are considered for the remaining quotas in essentially the same way. Again the more able men are apt to be sent to their first or second choices.

Program SKOOL seems to have accomplished goals set for it. More men than under Program OPERATE are selected for their first school recommendation--about 70 per cent get their first choice against 58 per cent with Program OPERATE; computer costs have been reduced to about 1/10th the cost of OPERATE; and, fewer men require hand assignment. Although the average payoff in SKOOL is slightly less than for OPERATE, we believe that the greater weight given to the recruit's choice of schools may compensate for this slight loss.

The last program in COMPASS is Program GENDET which assigns recruits who are not eligible for, or desirous of school training, to areas of general detail duty for on-the-job training in the Fleet. It is similar to Program SKOOL in that it considers each man's qualifications and preferences in making assignments.

We are presently attempting to evaluate the effectiveness of COMPASS. One study involving a group of 905 recruits, demonstrates the potential of computer assisted assignment. Using hand classification, 487 recruits who met school requirements were located--although the school quotas for that group were 545--a shortage of 58 men. The COMPASS Program took the same group of 905 men and was able to fill all school quotas--no shortages. Furthermore the test scores of the men assigned by COMPASS were slightly higher on the average than for conventionally assigned men.

Incidentally it has become increasingly difficult to demonstrate that COMPASS is significantly better than hand assignments, because hand assignment procedures have improved as a result of the work done while developing COMPASS. For example, the fact that filling quotas was emphasized and that assignment errors were found (and are costly) has lead to a general tightening up of all steps in the classification process.

The computerization of enlisted classification has many ramifications that we feel will have great payoff in the next few years. One aspect which I touched on earlier is in the more accurate prediction of individual performance. No longer are we bound by the usual methods of adding test scores (such as GCT + ARI = 110) in simple linear combinations. Now that all test scores are entering the computer routinely for COMPASS, the computation barrier has been removed and more refined prediction models may be used. At present we are investigating various ways of exploiting this potential.

One approach we are pursuing is in the use of moderator variables. A moderator variable is a variable, usually a test, but possibly something else like age, race, education, etc., which may by itself not be predictive of performance, but may improve the predictive power of other variables in our classification battery. It is conceivable that even though a certain test battery may predict recruit performance in, say, Electronics School with a validity of .40, a smaller subgroup of a sample may be predictable to the extent of perhaps .70. Our task is to devise ways of identifying the small subgroup and making the prediction program operational.

The computerization of enlisted classification has other welcome implications for future research. One of these is that we now have potential for redesigning our test batteries to maximize classification effectiveness whereas formerly only predictive effectiveness was being considered. Another advantage lies in the possibility of including measures other than aptitudinal and achievement ones, which in combination with the battery scores can enhance our accuracy of classification. As a matter of fact we are currently developing empirical achievement keys for a vocational interest inventory which appear promising in predicting school performance in combination with the aptitude measures. The inclusion of these non-intellective measures in the COMPASS programs could materially enhance classification.

Another advantage--this one will require a good deal of research before it can be exploited--is in the employment of decision theory in classification. Given the manpower availability and requirements which prevail today where should a recruit be assigned who has a 70 per cent chance of graduating from an Electronics School and 90 per cent chance of graduating from a Yeoman School? Decisions like this are being made daily on a very subjective and inconsistent basis. Techniques are being investigated for developing methods of scaling ratings on a common basis--for example criticality--to improve the quality of decision making. Another important area of investigation which is in progress is in the prediction of on-the-job performance. Currently, grades earned in basic level schools are the only available standard for judging the "success" of our predictions and though limited are accepted as primary criteria. Through additional research we believe it will become possible to develop specific measures of job success. Finally--and this last point is of great significance--the computerization of classification using improved criteria of performance will bring us closer to the goal of placing the individual in the job he's most suited for within the restrictions imposed by the Navy's needs. The ultimate success of any system of classification whether human or "non-human" must consider the individual and his satisfaction in the job and doing it well. Only when we are able to balance success, as viewed by the Navy, with success, as viewed by the individual can we consider our job accomplished.

ASSESSING HUMAN CAPABILITIES THROUGH EQUIPMENT-ORIENTED TESTS

L. R. Egan
Personnel Research Laboratory (Navy)
New Developments Research Department

Before I discuss our testing program, I think it would be appropriate to set the stage by briefly acquainting you with the work of my immediate organization, which in essence defines and delimits the scope of our testing.

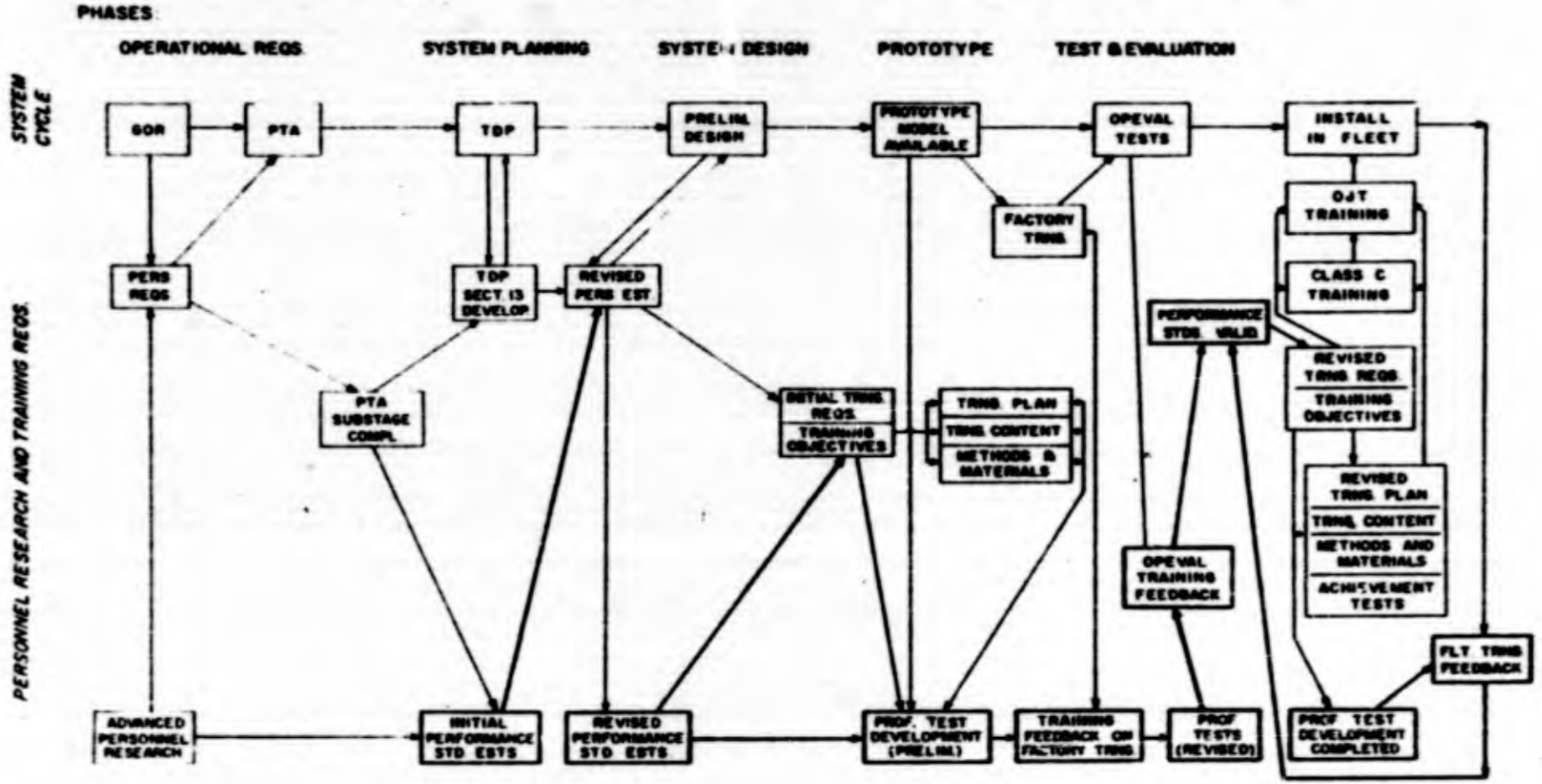
The New Developments Research Department is concerned with manpower requirements imposed by the introduction of new weapon and support systems into the Fleet. We are chartered to estimate, define, and evaluate the impact of new hardware systems along many human dimensions: for example, quantitative personnel estimates (a type of numbers game); functional task analyses (based on comparative man/machine functions as manifested at the earliest design stages); shipboard manning predictions; and training estimations. Not only are we committed to predict manpower and training requirements but also we attempt to verify these predictions as the hardware proceeds along the development cycle.

(Slide I) This slide illustrates the timing and documentation sequence that supports hardware and manpower requirements.

The first line is the development cycle of the hardware: from the GOR (General Operational Requirements) Stage, "the glimmer in the eye" of warfare specialists, to the final stage of Fleet introduction and use. As Human Factors specialists we tie into each phase, with varying degrees of effort, as diagrammed on the second line. You will notice that these cells reflect personnel and training estimates. They represent documentation, coordination, and applied research reporting in progressive degrees of specificity. The third line of this matrix is the evaluation phase, which runs concurrently with the other two phases. For example, human engineering and system effectiveness data are the bases for deriving evaluative instruments.

We are, therefore, interested in hardware requirements and the constraints and demands hardware places on the man in an operating environment. We want to know what he has to do (performance criteria) and how well he is doing it (evaluation). We as testing specialists begin planning for evaluation as early ir the development stage as do our human

PERSONNEL PERFORMANCE REQUIREMENTS AND TRAINING FEEDBACK IN THE RDT&E CYCLE



engineering specialists. (There is an overlay of responsibilities as the slide illustrates.) The latter specialists ask the question, "Will our Navy technician be able to perform the job?" and we, as evaluators ask, "How <u>well</u> will he be able to do it?" (End of Slide I)

We all realize, when we attempt to put our ideas into practice, that there is a considerable gap between a concept and its implementation. There are the everday agonies of not enough time, money, nor manpower: all the contingencies that constantly remind us of this finite world ("time for decisions and revisions which a minute will reverse," Eliot). Therefore, the attempt I have made to define our organization's role represents what we strive to do. We have had our share of successes and, of course, our burden of failures. I will speak to both.

One of the first systems which we tackled as a new organization (We were established less than 10 years ago.) was SUBROC, an anti-submarine weapon system, including both missile and fire control sub-systems. We prepared manpower estimates for this system early in its inception, at the GOR stage, at the evaluation level; however, we are still trying to assess its manpower impact.

With Promethean foresight, we realized that performance testing should be an integral part of our final evaluation program. Therefore, we derived a taxonomy of skills and concomitant knowledges that the SUBROC technician may be expected to encounter. This was in 1958-59. Not until recently, however, did we finally package the test booklets (1965-66). And facing the realities of the mercurial testing world, we are now repackaging them. What have we learned from this? First of all, this experience reinforced the expectation that testing is a painstaking, tedious, and gargantuan task; even more so, when it involves performance testing on equipment. We learned that for testing purposes we should have followed hardware development a little more closely. Changes in hardware specifications, if extensive enough, entail not merely test revision but an entirely new test. The Phoenix does not regenerate inself in the testing world; it must be reconstructed on a new site, cell by cell. We also found ourselves harrassed with the classic problems in test development: criterion determination, task sampling, scoring, validation, and the final demon, consumer acceptability

Because of the very nature of the tests we develop, mainly, job-sample performance, we can only in a limited sense use test technology which is based on a test's predicitive power.

AN/SKQ-1 TELEMETERING GROUND STATION EQUIPMENT

Tasks	Mission Contribution Rank	Sampling Guide	Test Section	Related Task(s)
1.0 SKQ-1 Preoperation Procedures				
1.1 Initial Setup of SKQ-1	1	Y		
1.2 Loading Tape Transport	2	A	2.3	
1.3 Loading Oscillograph Paper	3	Z		
1.4 Miss Distance Calibration	1	X		3.3, 3.6, 4.6, 611
2.0 SKQ-1 Operating Procedures				
2.1 Flight Recording	1	X		3.6
2.2 Direct Telemetric Recording	1	X		3.6
2.3 Direct Input Oscillograph Recording	7	X		3.6
2.4 Tape Playback and Oscillograph Recording	4	Z		3.6
2.5 Timing Signal Generation	7	Z		
2.6 Calculation of Miss Distance	5	Z		
2.7 Flight Record Analysis	5	Z		
3.0 SKQ-1 System Checkout				
3.1 Multicoupler and Test Cable Checkout and Measurement	1	X		3.3, 3.6, 4.4, 4.6,6.1
3.2 Power Supply Checks	1	Y		
3.3 Miss Distance Chassis Sensitivity Check	1	A	1.1	
3.4 Receiver Sensitivity Check	1	X		3.3, 3.6, 4.6, 6.4

Table 5. Job Requirements Summary, Sampling and Test Representation

AN/SPM-9 RADAR TEST SET

Chassis/Assembly	No. of* Malfunctions	Reliability Rating	Circuit Complexity Rating	Item Allocation
SPECTRUM ANALYZER		(1/5 x column 1)		(col.2 + col.3-1)
Video Preamplifier	26	5	2	6
Video Trigger Amp.	5	1	2	2
CRT Unit	10	2	5	6
IF Amplifier	2	1	3	3
Gate Generator	5	1	3	3
High Volt. Power Supply	13	3	1	3
Low Volt. Power Supply	25	5	4	8
SYNCHRONIZATION INDICATOR		(1/10 x column 1)		(col.2 + col. 3)
Trigger Amp. & Sweep Circuitry	37	4	5	9
Y Input Amp. & Pulse Stretcher	11	1	3	4
Y Driver Amp.	16	2	3	5
X Amplifier	2	1	1	2
Marker Generator	14	1	2	3
Axis Attenuator	1	1	1	2
Low Volt. Power Supply	34	3	3	6
High Volt. Power Supply	34	5	1	6
SIGNAL COMPARATOR	13	(1/5 x column 1) 3	5	(Col.2 + col.3-1) 7

*Based on Failure Reports accumulated at the U. S. Naval Ammunition Depot, Concord, California, covering the time period 1 January 1959 through 31 May 1962.

Figure 6
Test Item Allocations for Troubleshooting to Subassembly Subtests

As Glaser clarified (Psychological Principles in System Development, p. 421), we are involved in "criterion-referenced" not "normative-referenced" instruments. We do not correlate test results with future performance. We evaluate performance, in situ. We do not use test scores to order examinees along a continuum of relative effectiveness, in the normative sense, nor do we use these same standard scores to predict their performance. Therefore, a score of 90 on a proficiency (performance) test means something altogether different from a score of 90 on an aptitude test. If the aptitude test is reproted in percentiles, a student making 90 is assured that 90% of all examinees scored lower than he. With some statistical manipulations this individual's future performance can be estimated within a prescribed range of probability. However, if a Navy technician earns a score of 90% on a performance test measuring critical job skills, we would interpret this as meaning that he has not met the demands of the job. In evaluating his performance, we are not concerned at this point in time with what other technicians are doing but only with the fact that he is not contributing 100% to system effectiveness. In the case of extreme safety hazards, one omitted step or safety infringement constitutes total test failure.

In developing tests, we are mainly concerned with test criterion development and the demands of the hardware system. We begin with the analysis of system requirements, both man and machine, and then attempt to estimate their separate and interacting requirements. We derive our criterion data from hardware specifications and from system and equipment effectiveness reports. In the early stages of system development, we derive, from engineering data, functional statements of what the technician is required to do. Because the universe of tasks to be performed is usually extensive, task sampling is mandatory. As conditions warrant, a variety of performance criteria is employed: criticality, frequency, and complexity of task groupings. Redundancy across task groups is filtered out.

(Slide No. 2)

This is an example of task sampling on the SKQ-1 Telemetering Equipment. Tasks are listed in column one. Column two ranks, on a scale of 1 to 7, each particular task's contribution to the mission; these are in descending order; a rank of one indicates a task's direct function in receiving telemetering data. Column 3 is a secondary sampling, categorizing maintenance activities in terms of criticality and/or redundancy. Column 4 identifies subtest section. Column 5 cross indexes other related subtasks. When a task emerged from this sampling process, with the mission contribution rank/sampling combination of 1-A, B, C or 2-A, B, C, the task qualified for test representation. (End of Slide #2)

As a system proceeds downstream, there is expectedly more data available in the form of system effectiveness and equipment casualty reports. (Slide #3) This illustrates another methods of task sampling in which historical records of equipment operation (i.e., malfunctions) were used to ascribe weights to subtest areas. (End of Slide No. 3)

Out of the requirement for criterion-referenced tests, as well as the consideration of user acceptability, we have experimented with a variety of testing formats and designs. We have concentrated on the job-sample performance test in which the technician's overt behavior is observed and measured. We test him as he actually works at his job, either operating or maintaining equipment. We attempt to do what Lindquist (Preliminary Considerations in Objective Test Construction, p. 152), recommends:

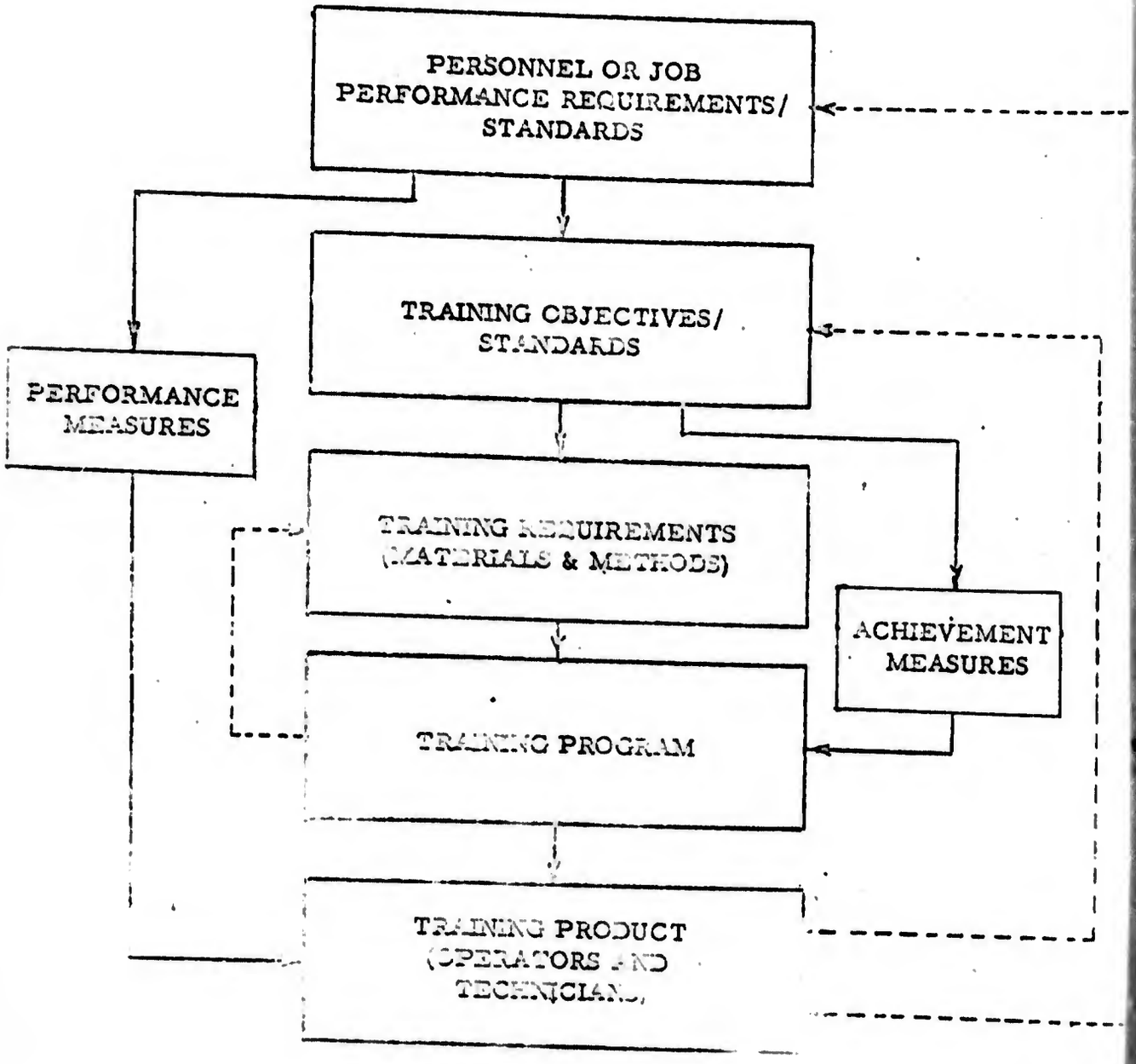
> ...the fundamental goal of the test constructor is to make the element of his test series as nearly equivalent to, or as much like, the elements of the criterion series as considerations of efficiency, comparability, economy and expediency will permit.

This quotation leads into the last demon I spoke of earlier, consumer acceptability. There are many constraints on test administration in an operating situation. We have been confronted with the majority of the, namely, time restrictions, equipment availability, or rather the lack of it, and the general reluctance of people to submit to examination. To make testing more palatable, without hopefully sacrificing validity and reliability, we have designed tests to fit the Procrustean bed of operational requirements.

Specifically, we have abbreviated administration time by constructing subtests to be administrered within a time frame of one hour; by devising symbolic (written) tests rather than performance; by providing alternate simulated problems ("assumed faults"); and by selectively sampling from among tasks within the total performance spectrum. These measures have markedly contributed to fleet acceptability because they minimize interference to shipboard routine.

What use is made of test results? As mentioned before, our proficiency tests are designed to be administered primarily aboard ship. They are used to detect weaknesses (and strengths) in the technician's ability to operate or maintain electronic equipments. They evaluate an individual's performance in terms of hardward requirements. Results from

IDEALIZED CONCEPT OF TRAINING DEVELOPMENT
AND EVALUATION PROCESS



tests, individually administered, are accumulated and analyzed in terms of general performance requirements. We can detect trends, for example, in a designated enlisted rating's ability to troubleshoot a particualr component within a particular system. Test results may show, for example, that of a sample of 15 FT's (Fire Control Technicians) tested, 10 failed to isolate troubles in the Signal Comparator section of the AN/SPM-9 test set. This information has specific meaning to those responsible either for shipboard training or for formal school programs. (Slide No. 4)

Our evaluation system is external to the school program. Our proficiency tests can be used in the school but do not necessarily measure a trainee's progress (or achievement) in school but rather his expected performance level on the job. If the school has defined its training objectives in terms of actual job requirements, then our proficiency tests do evaluate these objectives. In fact, our tests in many instances help training activities to more closely align their program with job requirements. (End of Slide No. 4)

As is the bane of researchers, we find ourselves left with more questions than answers to the measurement problem. Building proficiency tests based on equipment and system requirements is expensive and time-consuming. So we wonder, can we obtain the same information more expeditiously and with the same degree of confidence by using other techniques. We are now reassessing the validity of well-constructed survey instruments, couched in terms of performance requirements, and are evaluating their data-gathering power at various points along the hardware developmental cycle. We have found in our initial study, which employed proficiency tests, rating scales and survey instruments on the same sample of subjects, that there is a moderate amount of common variance in the data collected but not enough to convince us that tests do not have a unique function in the hardware-to-man assessment world. We are resharpening our tools and hope to apply more precision in evaluating our assumptions.

This is an overview of our testing program. For those of you who may be interested in more specifics concerning the rationales and formats we have developed for tests, I am distributing a sample list of our reports which are available from DDC.

tests, individually administered, are accumulated and analyzed in terms of general performance requirements. We can detect trends, for example, in a designated enlisted rating's ability to troubleshoot a particualr component within a particular system. Test results may show, for example, that of a sample of 15 FT's (Fire Control Technicians) tested, 10 failed to isolate troubles in the Signal Comparator section of the AN/SPM-9 test set. This information has specific meaning to those responsible either for shipboard training or for formal school programs. (Slide No. 4)

Our evaluation system is external to the school program. Our proficiency tests can be used in the school but do not necessarily measure a trainee's progress (or achievement) in school but rather his expected performance level on the job. If the school has defined its training objectives in terms of actual job requirements, then our proficiency tests do evaluate these objectives. In fact, our tests in many instances help training activities to more closely align their program with job requirements. (End of Slide No. 4)

As is the bane of researchers, we find ourselves left with more questions than answers to the measurement problem. Building proficiency tests based on equipment and system requirements is expensive and time-consuming. So we wonder, can we obtain the same information more expeditiously and with the same degree of confidence by using other techniques. We are now reassessing the validity of well-constructed survey instruments, couched in terms of performance requirements, and are evaluating their data-gathering power at various points along the hardware developmental cycle. We have found in our initial study, which employed proficiency tests, rating scales and survey instruments on the same sample of subjects, that there is a moderate amount of common variance in the data collected but not enough to convince us that tests do not have a unique function in the hardware-to-man assessment world. We are resharpening our tools and hope to apply more precision in evaluating our assumptions.

This is an overview of our testing program. For those of you who may be interested in more specifics concerning the rationales and formats we have developed for tests, I am distributing a sample list of our reports which are available from DDC.

Personnel Research Laboratory

Selected References

Development of a Job Sample Proficiency Test for AN/SPM-9 Radar Test Set Maintenance Technicians. Final Report, Bureau of Naval Personnel, Thiokol Chemical Corporation, Los Angeles 45, California, July 1963.

Proficiency Tests for Maintenance Personnel 6b1-6bw1 Talos Missiles AN/DSM-18b Test Set. Bureau of Naval Personnel, New Developments Research Branch Contract No. Nonr-3287(00), May 1962, (AD 422-924).

Proficiency Tests for Maintenance Personnel Sam Talos Missiles 6b1cw, 6bw1, and 6c1 AN/DSM-60 GMTS and AN/DSM-18c. GMTS. Technical Research Report, Bureau of Naval Personnel, May 1963, (AD 422-924).

Proficiency Tests for Missile Technicians Maintaining the AN/SKQ-1 Telemetering Ground Station Equipment. Volume III, Technical Report, Bureau of Naval Personnel, Dunlap and Associates, Inc., Santa Monica, California, February 1964 (AD 431-945).

Proficiency Tests for Missile Technicians Maintaining the BT-3, BT-3A and HT-3 Terrier Guided Missiles and the Mark 15 Mod 1 Tartar Guided Missile. Volume III, Technical Report, Bureau of Naval Personnel, Dunlap and Associates, Inc., Santa Monica, California, November 1963, (AD 422-503).

Proficiency Tests for Missile Technicians Maintaining the CM-122/DSM Signal Comparator. Bureau of Naval Personnel, Dunlap and Associates, Inc., Santa Monica, California, August 1962, (AD 476-503L).

Proficiency Tests for Missile Technicians Maintaining the Telemetry Ground Station Equipment AN/SKH-1 and AN/UKR-10. Technical Report, Bureau of Naval Personnel, Dunlap and Associates, Inc., February 1964, (AD 431-894).

Proficiency Tests for NTDS Maintenance Personnel: Computer Technician, System Technician. Final Technical Report, Bureau of Naval Personnel, Contract No. Nonr-3811(00), Matrix Report No. 63-7, July 1963, (AD 423-276).

Torpedo MK 44 Mod 1 Training Feedback Study, WRR 66-6, Bureau of Naval Personnel, Personnel Research Laboratory, April 1966 (AD 481-251L).

Training Feedback Information Requirements and Methods in the Research, Development, Test, and Evaluation of Navy Systems. Report No. ND 65-4, Bureau of Naval Personnel, Dunlap and Associates, Inc., Santa Monica, California, July 1964 (AD 608-247).

ULTIMATE USE OF COMPUTERS IN TEST INTERPRETATION

G. J. Anderson
National Computer Systems, Inc.
Minneapolis, Minnesota

Although manually scoring and processing test answer sheets is the basic technique used for obtaining the desired data, automated methods are generally much superior in all respects. The time factor involved is, quite obviously, disproportionately biased in favor of machine scoring. Overall accuracy and flexibility in the given measurements are also increased when machine methods are utilized.

Test answer sheets which have been designed for use with a scoring or reading device contain a fixed matrix of explicitly delineated discrete item response areas keyed to an associated question set. Although this does provide a form which is also quite amenable to manual scoring procedures, complexities in some test vehicles usually mandate a machine scoring be specified. A classic example of this is the Strong Vocational Interest Blank (SVIB). Much of the current development in automated scoring equipment could be traced directly to this test.

The first machines made to accomplish automated test scoring varied considerably in function and complexity. The desired processing functions dictated the totality of machine design, because all operations had to be built into the equipment hardware. Since the answer sheet format and physical size were tailored to machine requirements, the test itself was implicitly defined with respect to content and purpose.

Data provided by these early machines generally was the raw score values presented as a meter reading, or as a counter tally, to be so recorded by an operator. In some cases, a punched card was generated to allow a subsequent communication with a computer. Any extension of simple raw score value reporting, such as a raw to T score conversion process, meant that conversion tables would have to be stored in some sort of a memory device. Machine design would have to become even more complex in structure.

With the advent of readily available small-sized general purpose computers, a decision was made to return to fundamentals. This resulted in the construction of an answer sheet reading device to be used as a peripheral

equipment to the computer. This system works in all respects as a scoring machine, but possesses several important advantages:

a) Regardless of answer sheet or test design, full control is always maintained by the computer.
b) Answer sheet content may readily include such auxiliary data as age, sex, name, or various unique identification numbers, in addition to the usual question content.
c) Answer sheet design is quite variable in format, allowing at least some amount of "human-engineering" for a maximal ease in test administration.
d) All of the data, such as question item responses, or auxiliary data furnished by test content, may be retained on some data storage media to allow any type of subsequent data or item analysis to be implemented.

Using a computer with a peripheral document reading device as the basis for an automated test scoring system is becoming increasingly important to test users and test designers alike, because of the unprecedented data collection and retention possible. These features are currently being evaluated.

Any test scoring process first produces a set of scale values, which are used to quantitatively define the various parameters that the given test was designed to measure. Further processing of these data provides a considerable variety of currently realized output reports, ranging in content and form from an individual's profile report to complete group item or scoring statistics analysis.

All of these reports are purely factual in presentation, and as such, serve an extremely important and growing role in the entire testing field. Fortunately, or unfortunately, as the case may be, interpretation of these quantitative results may, in some cases, require either special training or experience, or an inordinate amount of time to accomplish.

The second step of this scoring process, therefore, especially with respect to individual data or profile reports, is to analyze these quantitative data and translate this to an actuarial description. This might be done in an implicit manner, such as a guidance counselor using an individual's SVIB profile to assist in directing attention to an indicated vocation; or explicitly, such as a clinical psychologist translating and reporting his analysis of an individual's MMPI test profile.

As is well-known and acknowledged, this second step in the scoring process - reduction of quantitative data to qualitative data - has very definitely not been subjected to automation. There are various reasons for this - duplication or simulation of the associative thinking modes necessary to implement this translation have not been clearly defined. An amazingly small amount of data is available on how men do solve problems, make decisions or perform general high-level intellectual funtions. Some terms for categorizing these activity types have been used, such as analyze, synthesize, organize or categorize, but close inspection of such terms and their definitions are, as yet, somewhat unsettling. There is also a problem, perhaps due to the disconcerting redundancy of the English language, in endeavoring to normalize a possible set of statements which could be used for a qualitative data report.

A computer is, in spite of the tremenduous technological evolution in both capability and capacity, a singularly unimaginative device. Except by sheer accident, it will do no more than programmed to do - in many cases, it will do considerably less. This means that everything which is to be accomplished with the aid of a computer must be very explicitly defined in all respects. The generation of quantitative data is immeasurable simple in this respect; the mathematics is well-known. Qualitative data generation is a relatively new concept, and much work and definition has yet to be accomplished in order to realize some degree of the full potentialities inherent in this application.

An example of a currently available test report containing both quantitative and qualitative data is the interpreted MMPI output as defined and developed by John S. Pearson, Ph.D. and Wendell M. Swenson, Ph.D., of the Mayo Clinic, Rochester, Minnesota. This report is a pioneering effort designed primarily to be used as a teaching or guidance device to assist anyone concerned with the various clinical aspects of behavior in the acquisition of the more subtle techniques of MMPI profile analysis. This particular report is medically oriented, and was designed primarily to be used by physicians in general.

The quantitative data portion of this report lists the raw and T score values of the cannot say, L, F, K and the ten clinical scales. Those raw and T scores which are subjected to K corrections are also listed, along with the appropriate T score scale plotting. The raw and T scores for the eleven research scales are reported, but not plotted.

The qualitative data portion of this report is comprised of a series of statements which conservatively

describe various implied aspects of the individual profile elevations. These statements, consisting of a series of succinct observations related to score elevations present, are generated during a three-step analysis of the individual score values.

The first step examines the variable set defined by the individual's age, sex and various clinical scale score values to generate a configural analysis of the given profile. This processing phase also tests the cannot say value for an indication of test completion, and tests the various T score values for the number of primed scores.

The second step in this data generation is to sort the L, F and K scales by magnitude and generate the appropriate comment output. Finally, the remaining clinical scale values are sorted by magnitude, and the statement related to the degree of elevation for the given scale is generated and printed out.

It is to be stressed that this first application of qualitative data processing by a digital computer is quite modest in intent and specific in nature. While the statements generated are entirely correct within their framework of reference, a specific user should tend to augment these statements with implied interpretations.

A continual expansion in the use of digital computers can only result in the extension and refinement of techniques that are currently used for quantitative data processing. This is markedly evident in some of the various aspects of item analysis - question content, number of questions used, or question area overlap may well be derived by the use of a computer. Processing studies are currently being made of this type of analysis, and consequent developments would certainly implement the capability of automated test design.

It is in the area of definition of qualitative data analysis that the greatest changes will be effected. Problems of term definition, such as previously stated, are being examined by a number of people working in divergent fields of endeavor. The totality of theory developed from this work, used in conjunction with possibilities afforded by the newer concepts in computer organization and control, will eventually result in a very real convergence of anticipated results toward what is now termed an ideal.

A most sincere acknowledgement is given to John S. Pearson, Ph.D. and Wendell M. Swenson, Ph.D., of the Mayo

Clinic, Rochester, Minnesota. Their combined research efforts made possible the implementation of the interpreted MMPI output described.

BLANK PAGE

A QUESTIONNAIRE TO IMPROVE A TESTING PROGRAM

J. E. Partington
US Army Enlisted Evaluation Center, Fort Benjamin Harrison
Indianapolis, Indiana

One of the Army's big testing programs is a part of the Enlisted Evaluation System. This system affects every aspect of a soldier's career - his assignment, his utilization, his promotion, his training, his pay, and his retention in the Army. It is a means of determining annually by means of a test and a rating scale how much a soldier knows about his job (his Military Occupational Specialty or MOS) and how well he performs in his MOS. It provides an objective and comparative measure of job competence for eligible soldiers in the same MOS skill level and pay grade.

The Enlisted Evaluation System gives each soldier individual attention. Job evaluation provides for him both motivation and interest. Complete knowledge of job requirements furnish him with a sense of accomplishment and a comparative evaluation of his job proficiency among his peers which provides for him a competitive challenge. Objective results of the Enlisted Evaluation System are used in connection with promotions, proficiency pay, retention, assignments, and utilization of enlisted personnel.

The MOS evaluation cycle begins for the soldier when he receives his test aid about three months before taking the paper-and-pencil test. The test aid serves as a study reference and a training aid. It contains suggestions on how to prepare for the test, a list of study references, non-resident Army school courses, and job requirements within the major subject-matter areas of the MOS skill level.

Two measuring devices are used to evaluate the job knowledge and job proficiency of the soldier - the Commanders Evaluation Report (CER) and the MOS evaluation test including a performance test when appropriate. The CER is a rating scale composed of 12 factors which cover personal characteristics of the soldier. The MOS evaluation test is composed of 125 multiple-choice items (questions) which are concerned with job problems and job situations. The score on the test, the CER, and the performance test (when administered) are statistically combined to obtain a raw composite score. This score is then converted directly to an MOS evaluation score using the scale of the Army standard scoring system. The evaluation score is the one reported to the soldier and to his commander. Raw composite scores of soldiers in the same MOS skill level and pay grade are grouped together. The minimum score for MOS qualification and for promotion are determined. MOS evaluation scores ranging from 40 to 160 are then developed for the group. A soldier must obtain an evaluation score

of 70 or higher in order to verify his MOS. Soldiers who obtain an evaluation score of 110 or higher are considered promotion qualified.

(EDR VU-GRAPH) The MOS evaluation cycle ends for the soldier when he receives his MOS Evaluation Data Report (EDR). This is the "soldier's report card" and gives the results of his MOS evaluation. In addition to showing the soldier's evaluation score, the EDR shows a profile of his strengths and weaknesses in each broad subject-matter area of his MOS skill level based on his relative standing in each broad subject-matter area of his test. This information identifies for the soldier those areas of his MOS in which he needs additional study and training. If he strives to improve those areas in which he received "low" or "very low" on his EDR Profile, he should become a more effective soldier thus giving him a sense of accomplishment. He should also improve his MOS evaluation score. This is especially desirable if he failed to verify his MOS or failed to qualify for promotion, or for proficiency pay.

The success of the MOS evaluation testing program depends not only on the development and use of valid and reliable measuring instruments, their proper administration and scoring but also on (1) the use which the soldier and his commander make of the results of the program as shown on the EDR for the soldier's improvement; (2) the success the soldier has in obtaining reference materials necessary for test preparation; (3) the timeliness with which the soldier receives the test aid; (4) acceptance by the soldier of the guidelines in the test aid; (5) and the extent to which study references helped the soldier in obtaining information about equipment pertinent to the MOS but with which he is not familiar. Acceptance of the MOS testing program and use of its results by the soldier is absolutely essential to its continuing success. In order to obtain information about these important areas, a questionnaire was developed for administration to a sample of soldiers immediately after administration of MOS evaluation tests.

The MOS Evaluation Test and Test Aid Questionnaire were designed to elicit frank and anonymous opinions about the MOS evaluation tests, test aids, availability of reference materials, the Evaluation Data Report (EDR), study habits, and other related aspects of the Enlisted Evaluation System. A questionnaire composed of 21 multiple-choice questions was administered to a random sample of soldiers during the November 1965 and February 1966 test periods from which almost 5,000 useable answer sheets were obtained.

Questions are a means of communication and communicating just what we want to say is really not an easy task. As far back as 1936, a study indicated that in the field of commercial

research, three experts in every four were critical of improperly worded questionnaires. The person who asks a question thinks that it is clear and will be understood. To each of us our statements seem clear. To the listener, however, it may be a different story. A fellow in an office told recently of a household incident of which he had been an innocent but perplexed spectator. Our friend had called a Venetian Blind repairman to come pick up a faulty blind. The next morning, while the family was at breakfast, the doorbell rang. Our friend's wife went to the door and the man outside said "I'm here for the Venetian Blind." Excusing herself in a preoccupied way, the wife went to the kitchen took a dollar from the food money, and pressed it into the repairman's hand. She gently closed the door and went back to the breakfast table. "Somebody collecting," she explained as she poured the coffee.

I don't think that the material in our questionnaire has been misunderstood. We did, however, revise the questionnaire and expanded it to 41 questions for administration to soldiers in May 1966 and August 1966 test periods. The results from the November 1965 and February 1966 questionnaire were very similar to those obtained on the revised questionnaire administered in May 1966. This paper is concerned with the results obtained from the revised questionnaire administered in May 1966 from which 2,878 useable answer sheets were obtained.

The questionnaire was concerned with soldier's opinions in six major areas as shown on this Vu-Graph:

(USE VU-GRAPH)

1. Evaluation Data Report.
2. Test questions on unfamiliar equipment.
3. Test Aid.
4. Individual or group study.
5. Reference material.
6. Preparation for tests.

I will now present the separate items of the questionnaire with percent of responses followed by implications for improvement of the testing program.

Item 1. If you received an Evaluation Date Report (EDR) from last year's test, was it helpful to you in determining those subject-matter areas, if any, for which you needed additional study to prepare for the test you just completed?

A.	very much	35%
B.	some	40%
C.	very little	11%
D.	not at all	14%

The EDR is favorably accepted by a majority of soldiers. 75% said that it was helpful in determining areas in need of additional study relative to the test which was just completed.

Item 2. Answer this question ONLY if you received an EDR last year and if the EDR profile indicated low or very low in any subject-matter areas. Did your immediate supervisor or commanding officer encourage you to study in those areas?

A.	yes	34%
B.	no	66%

Two-thirds of the soldiers said that their commanders did not encourage them to study in the areas of weakness.

Item 3. In the test you just completed, approximately how many questions referred to types of equipment with which you were not familiar?

A.	more than 25	29%
B.	11 to 25	21%
C.	1 to 10	32%
D.	none	18%

One-half of those answering this item said that there were 25 or more test questions on unfamiliar equipment.

Item 4. If the test contained questions on equipment not familiar to you, how many of them, if any, could you answer because you had worked on or had been trained on a similar (or substitute) type of equipment?

A.	most of them	23%
B.	few of them	61%
C.	none of them	16%

84% did obtain some benefit from having worked on or having been trained on similar type equipment.

Item 5. If the test contained questions on equipment NOT familiar to you, how many of them, if any could you answer because you had studied the references listed in the test aid for that type of equipment?

A.	most of them	17%
B.	few of them	45%
C.	none	11%
D.	I did not study the references.	27%

62% did obtain some benefit from having studied the equipment references listed in the test aid.

Item 6. Did you receive a test aid for the current test administration period?

A. Yes, through normal distribution channels	66%
B. Yes, through a source other than normal distribution channels	6%
C. I did not receive a test aid	28%

It was disturbing to note that 28% stated that they did not receive a test aid at all. An analysis of the replies on the February 1966 questionnaire of three geographical areas concerning receipt of test aids revealed that 41%, 47%, and 49% respectively stated that a test aid was not received. It was apparent that more information concerning timely receipt of test aids is needed and some remedial action is necessary.

Item 7. Approximately how far ahead of the test date were you informed that a test aid was available for you through normal distribution channels - regardless of whether or not you actually received it?

A. 2 months or more	41%
B. at least one month but less than 2 months	24%
C. less than 1 month	15%
D. at no time	20%

Item 8. Answer this question ONLY if you received a test aid. Approximately how far ahead of the test date did you first have the test aid in your possession?

A. 3 months or more	16%
B. at least 2 months but less than 3 months	40%
C. at least 1 month but less than 2 months	32%
D. less than 1 month	12%

Answers to this item indicate that most soldiers actually obtained a test aid at the same time they were notified of its availability.

Items 9 through 14.

How many hours of individual and/or group study, if any, did you spend in studying for your MOS evaluation test?

9.	A.	over 50 hours	12%
10.	A.	26 thru 50 hours	10%
11.	A.	11 thru 25 hours	19%
12.	A.	6 thru 10 hours	18%
13.	A.	1 thru 5 hours	16%
14.	A.	none	25%

75% spend one or more hours studying for their test with 12% spending more than 50 hours. The time spent, however, varies considerably among MOS Codes. Of 16 MOS Codes for which sufficient replies were available, for February 66 and 10 MOS Codes for November 65 the range for no time spent studying was from 8% to 40% and 2% to 39% respectively. Note that 25% spent no time studying.

Item 15. How far ahead of your MOS Evaluation Test administration date did you begin to study for your test?

A.	2 months or more	24%
B.	at least 1 month but less than 2 months	32%
C.	at least 2 weeks but less than 1 month	20%
D.	less than 2 weeks	24%

Over one-half of the soldiers began studying at least one month or more before their test.

Item 16. How useful was the test aid to you in preparing for the test?

A.	much use	27%
B.	some use	66%
C.	no use	37%

93% found the test aid useful.

Item 17. Of the references you wanted from those listed in your test aid, how many were you able to obtain?

A.	all of them	14%
B.	at least half of them	45%
C.	less than half of them	36%
D.	none	5%

It appears that reference availability is a problem of varying degrees for 86% of the soldiers responding.

Item 18. How much difficulty did you have in obtaining the references you wanted from those listed in your test aid?

A.	no difficulty	23%
B.	some difficulty	46%
C.	much difficulty	26%
D.	I did not try to get the references	5%

The results on this item are consistent with the previous one. It appears that some improvement in reference availability is needed.

Items 19 through 27. From what source or sources did you obtain study reference material listed in your test aid?

19.	A.	personal reference material	54%
20.	A.	unit reference file	51%
21.	A.	unit commander or first sergeant	11%
22.	A.	post library	14%
23.	A.	education officer (education advisor)	9%
24.	A.	borrowed from another person	40%
25.	A.	manuals which accompany equipment	25%
26.	A.	other	20%
27.	A.	I did not try to obtain any references	2%

Item 28. If you were unable to obtain the references you wanted, did you bring the matter to the attention of your immediate supervisor?

A.	Yes	35%
B.	No	35%
C.	I was able to obtain the references I wanted	30%

One-third did not bring their reference problem to the attention of their supervisors.

Item 29. Which of the following statements best describes the amount of reference material listed in your test aid?

A.	too much material for the time I had to study	46%
B.	about the right amount of material for the time I had to study	30%
C.	I would have studied more material	24%

It is interesting to note that 24% stated that they could have studied more reference material.

Item 20. (Feb 66) Was the listing of specific chapters or sections of references in your test aid useful to you in studying?

A.	yes	83%
B.	no	17%

Items 20 and 21 from the questionnaire administered in February 1966 are inserted here since they do not appear in the questionnaire administered in May 1966. Previously, specific chapters or sections of references were not given, thus requiring the soldier to find the appropriate parts of the references to read.

Item 21. (Feb 66) Study references are now listed in the test aid by subject-matter areas. Did you find this method of listing references helpful?

A.	yes	86%
B.	no	14%

Previously references were listed at the end of the MOS requirements, each reference listed once and its applicability to skill level checked in the skill level column. It is obvious that the changes in the Test Aid format for references has been overwhelmingly accepted by the soldier.

Item 30. Did you use any Army Extension Course(s) or Subcourse(s) to prepare for your MOS evaluation test?

A.	yes	13%
B.	no	87%

Note the large percent (87%) who did not use an extension course.

Item 31. Approximately how much of the extension course study material did you complete?

A.	all of it	24%
B.	two-thirds or more	28%
C.	one-third but less than two-thirds	25%
D.	less than one-third	18%

52% finished all or most of the extension course.

Item 32. Did the Army Extension Course(s) or Subcourse(s) you used <u>help</u> you to prepare for your MOS evaluation test?

A.	very much	34%
B.	some	47%
C.	very little	18%

81% did get help from extension courses taken.

Item 33. Did you attend any training sessions organized by an Education Officer or Education Advisor to help you prepare for your MOS Evaluation Test?

A.	yes	4%
B.	no	96%

Note the large percent who did not attend Education Officer training courses. Perhaps many of the Education Officers did not offer any training courses.

Item 34. Approximately how many of the training sessions, organized by an Education Officer or Education Advisor did you attend?

A.	21 or more	12%
B.	11 thru 20	15%
C.	6 thru 10	25%
D.	1 thru 5	48%

Item 35. Did the training sessions, which you attended and which were organized by an Education Officer or Education Advisor, help you to prepare for your MOS evaluation test?

A.	very much	27%
B.	some	51%
C.	very little	22%

78% obtained help from Education Officer training courses.

Item 36. Did you attend any training sessions organized by your Unit Commander to help you prepare for your MOS evaluation test?

A.	yes	7%
B.	no	93%

Note the large percent who did not attend Unit Commander training sessions. Training sessions may not have been available.

Item 37. Approximately how many of the training sessions, organized by your Unit Commander, did you attend?

A.	21 or more	7%
B.	11 thru 20	16%
C.	6 thru 10	26%
D.	1 thru 5	51%

Item 38. Did the training sessions, which you attended and which were organized by your Unit Commander, help you to prepare for your MOS evaluation test?

A.	very much	24%
B.	some	52%
C.	very little	25%

76% did get help from Unit Commander training sessions.

Item 39. Did you attend any informal group study sessions organized by you and/or others in your MOS to help you prepare for your MOS evaluation test?

A.	yes	22%
B.	no	78%

Note the large percent who did not attend informal study sessions. Such sessions were probably not organized in many places.

Item 40. Approximately how many of the informal group study sessions, organized by you and/or others in your MOS, did you attend?

A.	21 or more	5%
B.	11 thru 20	10%
C.	6 thru 10	29%
D.	1 thru 5	56%

Item 41. Did the informal group study sessions, which you attended and which were organized by you and/or others in your MOS help you to prepare for your MOS evaluation test?

A.	very much	34%
B.	some	54%
C.	very little	12%

88% obtained help from informal study sessions.

The following conclusions are based on results of the questionnaire along with recommended actions:

a. A substantial number of enlisted personnel (28%) stated that they did not get a test aid. A large number (84%) stated that they did not get a test aid a full three months ahead of the testing date. An MOS evaluation test aid availability questionnaire has been sent to selected Unit Commanders who are responsible for distribution of test aids to soldiers. Those selected were in the areas in which

the largest percentage of soldiers stated that they received their test aids less than 1 1/1 months before the test period. We want to find out when the company commander received his supply of test aids, to what percent of the soldiers he was able to give a personal copy of a test aid, and what method he uses for their distribution. Answers to these questions will help in bringing about a situation which will assure that every soldier receives or at least has available to him a test aid for use in preparing for his test.

The Enlisted Evaluation Center has taken effective action to assure that test aids get distributed to Test Control Officers in time to have them distributed to soldiers 90 days prior to the first day of the test administration month. Expeditious distribution after leaving the TCO does not always seem to occur, however.

b. The Evaluation Data Report is a useful instrument in helping enlisted personnel to identify areas of their MOS in need of improvement. About 70% answered in the affirmative. Studies under way to determine if results of this report can be used to identify training needs in small units. Unit Commanders will be encouraged to use the EEDR.

c. The test aid is a useful instrument to enlisted personnel in preparing for testing. 93% answered in the affirmative. There is overwhelming acceptance of recent innovations for listing references in test aids. About 85% answered in the affirmative. Test aid format will remain very much as it is now.

d. There seems to be a desirable balance between those who think there are too many references (46%) and those who could have studied more (24%). 30% said the number of references was just right.

e. Many enlisted personnel (54%) have personal reference material. Recommendations will be made to (1) encourage all Army installations to use "interlibrary" service and AUTOVON to help obtain technical publications which seem to be hard to get; (2) ask Publication Centers to review their policy on cut-off dates for publications requisitions which are given a status of "Due-Out"; (3) encourage enlisted personnel to cooperate in keeping various publications' reference files and libraries intact by not removing references for their personal files and to return promptly reference material for which they have signed; (4) encourage Unit Commanders and Education Officers to cooperate and assist in providing needed study references.

f. 87% stated that they did not use Army extension courses or sub-courses in preparing for their test.

Appropriate courses and subcourses are listed in each test aid. A recommendation will be made that the appropriate agency give additional information Army-wide concerning the availability and usefulness of extension courses and subcourses.

g. 96% stated that they did not attend training sessions organized by the Education Officer. A recommendation will be made to the appropriate agency that there is a need for more emphasis on Education Officer organized training sessions to assist soldiers in preparing for their tests.

h. 93% stated that they did not attend any training sessions organized by the Unit Commander. A recommendation that the Unit Commander give more attention to MOS training will be made.

i. 78% stated that they did not attend any informal group study sessions. Such sessions will be encouraged by recommendations to appropriate personnel and through appropriate publications media.

We are convinced that information about and the importance of the Enlisted Evaluation System will receive even greater attention at all levels of command. The system is already a success but attention to some of the areas highlighted by the results from our questionnaire will make the system even more successful. This is a good time for me to recall a story that happened on a golf course and follow the advice given One golfer said to the other, "The traps on this course are very annoying, aren't they?"

The other golfer, who was trying to putt, replied, "Yes, they are. Would you mind closing yours?"

THE POSITION OF ARMY MOS PROFICIENCY TESTING IN AN OPERATIONS RESEARCH CONTEXT

Raymond O. Waldkoetter
Evaluation and Analysis Branch
U.S. Army Enlisted Evaluation Center

In developing a test to measure a military occupational specialty (MOS), the approach has sometimes been to view the task as a limited interaction of testing stimulus, soldier, and the responses, with some filtered information processed out for certain selected measurement standards. Somehow this conceptual treatment has maintained its adequacy, but may have impeded the broader application of measurement data toward the solution of complex personnel selection problems. The optimum utilization of Army proficiency testing data can best be achieved, in my opinion, as the Enlisted Evaluation System is placed in the context of operations research (OR).

Since the Second World War's scientific and operational developments, when the British became exposed to very heavy operational requirements and had to design a comprehensive technique for rather immediate decision-making needs, the applications of the OR methodology have accelerated under the pressure of the growing vastness of difficult situations, demanding solutions based upon something besides experienced military judgment. The simple explanation of OR for the purpose of the present paper may be stated as that strategy, involving the use of mathematics and statistics, in deriving models or equations for operational activities which make up the subordinate elements of an organization thereby controlling the effectiveness of that total organization.

The point may be taken as to how this approach uniquely differs from a Systems or Cybernetics outlook. Frankly, I really cannot say in this short space with any conclusive evidence. The similarities most likely would surpass the isolated differences. But to make an attempt at contrasting, I would hazard that OR can encompass the structural implication of the term Systems and the functional emphasis reflected by Cybernetics.

The OR concept regarding Army proficiency testing makes a systems analysis of the testing process, production, and administration essential, with the identification of the various possible uses for test data. Within the adequately constructed measuring instrument one should find and be able to relate data covering the essential job elements, skills, and training objectives necessary for acceptable job performance. From this information, data can be applied to those situations where the statistical value of test information can be utilized to predict the outcome of some decision, that is inher-

ently related to a statistical model favorably adapted to a personnel measurement problem.

The decision-making processes in personnel management may then be aided by using much under-developed data being stored in the wide range of something of around 800 Army proficiency tests. The organizational structure must be regarded more dynamically, in order that the collected test data are not forced into a narrow channel of short-term administrative support. Generally, an outlook and program, which interrelates proficiency testing into an occupational evaluation system or data retrieval complex, can more efficiently use and disseminate the testing data that is appropriate for more than simply verifying the satisfactory assignment of a designated MOS.

Let us not pursue a too optimistic course here in assuming test data will fill some as yet uncharted personnel requirement, but still the possibilities are yet mainly untapped in bringing the position of proficiency testing well within a systematic program of operations research.

Creation of an OR team, that signifies a group as well as a procedural system, can do the structuring which will place the proficiency testing program in a sequential order whereby it is capable of exchanging information with other personnel systems. The definitive steps required to effect this positioning consists largely of clearly diagraming and relating the testing function to all the given dimensions of the enlisted occupational structures. The job description can profit from testing research in which skills may be validly identified; training courses may gain through analyzing pertinent test results that validly indicate at what level certain terminal behaviors must be, to predict some degree of acceptable proficiency; and, in assignment of personnel, test data may contribute to the statistical applications where assignment is determined by various alternative skill profiles to maximize the use of available personnel. Such relationships and the resultant store of data can be repeatedly analyzed, updated, and transmitted to numerous users who can contribute or inquire to help make the operational use of the test data as responsive as possible to a multitude of personnel research questions.

Posing a question, momentarily, as to the worth of the operations research context, wherein the better application of testing results may be hoped for, I must express a strong affirmation for the techniques being marshalled to provide the initiation of an expanded operational test data usage. While objectives in setting up a highly integrated proficiency testing system take on a more comprehensive character, they are not to be perceived as more difficult to define. With the diligent use of the method known by the title, Program Evaluation and Review Technique, or more familiarly as "PERT,"

the best probable visualization of a program's plans, objectives, etc., is capable of being diagramed with the end goals duly established for common agreement and convergence.

A contextual viewpoint that positions proficiency testing, so that its objectives are made to specifically overlap with other related personnel systems, gives in turn added reliability to the selected objectives. The products in such a context should then give the kind of output that is designed to fall within designated limits of expected performance or operation.

Through the experimental involvement of operations research to encourage the wider assimilation and redirection of test data, a number of investigations surely have to be executed before the methodology will have a valid or satisfactory set of products. Exploring the applications of operations research will undoubtedly lead into the creation of probable situations, having contending courses of action, with only one course of action yielding the greatest payoff. Some sort of simulation seems appropriate to delineate the effects testing produces, and the interchangeable data sources generated by a controlled system of reinforcing components of occupational information.

When reciprocating systems are all tied into a higher level system, where the objective is so projected to classify the various dimensions of data support and emphasis for management decisions, the outcome must succeed in giving a workable solution that will account for the principal situational variables. As the optimum state of resolving proficiency testing problems and other related personnel requests unfolds, then newly constructed personnel research models are brought within the arsenal of advancing operations research designs.

Upon the knowledge of imminent progress in this field, it would appear personnel research can only acquire certain kinds of information via the Army proficiency tests. Looking out over the field of Human Factors research, I would like to express a few more of my own limited observations.

Personnel investigations are adapting to the operations research technique in this area. Last May I was able to obtain an invitation to a MORS conference (Military Operations Research Symposium) at the Naval Post-Graduate School, Monterey, California. There the Human Factors working group undertook to discuss applications of OR to personnel decisions. After brief exposure to the MORS conference it became obvious to me that assessment of individual and group job performance was extremely critical to the concepts that were being described and entertained for research possibilities. This extension of Human Factors research which feeds on individual and group

job performance measures can be dependable just as long as the test data are accurately portrayed and validated.

The trend and compelling attraction for using operations research methodology within our testing program will lead to the betterment of the testing process, and the profitable exchange with other related components of an occupational evaluation and information data system which is built on maximum computer usage.

Currently, the Enlisted Evaluation Center is eagerly watching events, through the Enlisted Personnel Directorate, at the planning level in the office of Personnel Operations, our principal headquarters. There a Statistical Management Unit is beginning to take shape and in coordination with the Personnel Data Systems Office will oversee and possibly control the conduct of certain key research projects with the OR flavor. Such studies as an Optimized Enlisted Assignment System, Multiple Regression Equations Used to Assign Individual Training, the Occupational Information Data Bank, and Centralized Test Scoring will have highly useful consequences for the Enlisted Evaluation System in the months and immediate years ahead. However, the progress of these studies could be limited in part or expanded according to developments we have made and will make.

A system is not an "island unto itself" and neither is it totally dependent upon other activities. With the proper integration of test theory, computer science, well thought-out studies of test simulation, and worthwhile inquiries for data from DDC (Defense Documentation Center), the Enlisted Evaluation System should materially expand and prosper in being of "service."

As we have acquired new insights and friendships here in the Military Testing Association, and at corresponding meetings on Military Psychology (Division 19) in the American Psychological Association, we at the Enlisted Evaluation Center, and our fellow testing organizations, should only anticipate many new horizons and uses for testing data and services. There should be a good deal of human appraisal going on, with the continuous cooperation between our military and civilian staffs, in the endeavor to build better testing measures while using the reference points of critical insight and professional standards.

APPLICATION AND UTILIZATION OF CUTOFF SCORING PROCEDURES

MILITARY TESTING ASSOCIATION PRESENTATION

Mr. Dale Baker
US Army Enlisted Evaluation Center
Evaluation and Analysis Branch

The theories expounded to explain human appraisal are many and varied. We are proposing no new theories, but we have tried to make our appraisals of the enlisted man realistic and valid. In order to arrive at an equitable solution for the various problems encountered in our procedure, we have developed a cutoff scoring system. This is exactly as it sounds, a system to categorize enlisted men in relation to their peer groups. We feel we have developed sufficient categories, and evolved scoring systems complex enough, yet simple enough to provide the Army with good sound management tools.

Of paramount importance in developing our system, we felt it necessary to give the Army leeway to use some of the fine managerial practices developed through years of personnel testing. Yet in all fairness to the individual soldier trying to attain certain goals, we had to limit the perogative of the commanders in certain areas. We have tried to limit the effect of personal feelings and bias in the overall evaluation of a soldier.

Many critics of personnel testing contend that tests are poor selection tools. We wish to stress the fact that in many cases tests are the most valid source of selection data available. Since our instruments are sufficiently valid, we have found them a good method of evaluating the individual enlisted man. A good deal of my presentation today will deal with the practical application of these measuring instruments.

By giving individual commanders valid instruments with which to manage, direct, and train personnel, we feel we have greatly enhanced the ability of the Army to deal effectively with the problems of promotion, verification, classification, assignment, and so forth.

Three instruments are used to evaluate the enlisted men. These include the MOS evaluation tests (ET) which are developed to test the job knowledge and technical ability in most MOS. Further, they offer an objective method of determining the necessary knowledge and abilities of the men. A second instrument is the commander's evaluation report (CER). The CER is intended to provide an assessment of the soldier's job performance and potential for advancement by his two immediate supervisors. While all types of ratings assume that the rater <u>can</u> and <u>will</u> rate the soldier fairly and accurately, ratings often reflect the subjective biases and personal feelings of the

rater, and thus, are not perfect. They are, however, the best available methods of evaluating a large number of soldiers performing duties for which no objective measure can be obtained. Correlations between the MOS evaluation tests and the CER are held at as low a level as possible to insure independent measures. A third instrument used for evaluating enlisted men is the performance test (PT). These are of limited application because they have been developed for only a few MOS. The tests which have been developed include Bandsmen, Intermediate Speed Radio Operator, Radio Teletype Operator, Court Reporter, Stenographer, Clerk Typist and Interrogator.

A large majority of the EM are evaluated using a combination of these instruments. The most common evaluation is based on a weighted combination of the MOS evaluation test and the CER. Where the performance test is applicable this is also combined. Two methods of weighting the components are used. The first system is a formula of established weights in which the objective evaluation test and performance test receive proportionately greater weight than the subjective commander's evaluation report. These proportions are set on the raw scores, and the resultant raw composite score is the actual basis of the cutoff scoring system now employed. The second system is derived during the validation studies currently being done on each MOS. These weights are the beta weights which indicate the optimal weights each component should contribute to attain the highest multiple correlation. Validity coefficients for each system of weighting are calculated in each validity study. The two systems generally yield very similar results.

The final step in obtaining a score from which evaluations can be made is a linear transformation of the raw composite score to an evaluation score (ES). Simply stated the scores from the evaluation test and the commander's evaluation report and the performance test, when applicable, are combined into a composite score and this score is transformed into a corresponding score on the Army's standard scale. This does not change the relative values of the composite scores but converts them to a forced distribution based on parameters developed using the raw composite score. Using this forced distribution, on which the cutoff scores are set, promotion, verification, and nonverification are determined.

Evaluation scores are those scores which denote a soldier's knowledge, skills, and ability to perform particular MOS duties. These scores range from a minimum of 40 to a maximum of 160. In order to be considered qualified in his MOS skill level and pay grade, a soldier must make a score of at least 70 to verify or denote his performance is minimally acceptable. A score of 100 means that the soldier is about average for the skill level and pay grade of the MOS. A score

of 110 or more places a soldier in a position for promotion and indicates the upper 1/3 of the group.

In order to best explain our system of scoring and the cutoff method we use in the program, it will be easier to begin at the bottom of the scale and work up.

A new concept for the Enlisted Evaluation System, known as the "Minimum Hurdle" was implemented in the August 1965 Test Period. This concept established an absolute minimum score for each of the evaluative instruments used in the USAEES. The minimum absolute score has been established by DA at one chance standard deviation above the chance mean. This agrees with Gulliksen's recommendation that in order to signify any knowledge of the examination subject matter, the chance score must be exceeded by one to two chance standard deviations.

The minimum hurdle concept is designed to eliminate the "compensation effect" that is, a high score on one evaluative instrument compensating for a low score on another instrument. Separate absolute minimum cutoff scores are established for each measuring device. If an EM does not exceed each of these "minimum hurdles (scores)," he is classified as unqualified and receives an ES of 40. Use of the multiple or "minimum hurdle" is based on the consideration that an EM needs at least a minimum amount of "X" (MOS technical knowledge, as measured by the MOS evaluation test), a minimum amount of "Y" (personal qualifications and job performance, as measured by the CER) and when appropriate, a minimum amount of "Z" (specific job performance proficiency as measured by the MOS performance test) to be MOS qualified. No amount of "X," "Y," or "Z" can compensate for one or more of the other measures below the required minimum.

In the past, a soldier who maintained an excellent working relationship with his supervisor and received a high CER score, yet demonstrated little mastery of his MOS when tested, could still score above the minimum evaluation score (70) required for MOS verification. Likewise, the soldier with excellent knowledge of his MOS as indicated by his evaluation test, who received a very low CER score, could meet or exceed the minimum MOS verification score.

Under the present system, a soldier who fails to exceed the minimum absolute score on any one of the evaluative instruments will be declared unqualified and will receive an evaluation score of 40 regardless of his score(s) on other instrument(s). A soldier can no longer verify his MOS by compensating for a low score by an extremely high score on another instrument.

The second cutoff set in this system is the minimum verification score. An evaluation score of 70 is necessary for an

EM to verify his MOS. Any score between 40 and 69 indicates insufficient job knowledge, technical proficiency, or personal ability to be proficient in the MOS. An initial verification raw score is recommended by the experts in the MOS who are responsible for writing the MOS evaluation test. These recommendations are given to USAEEC to examine, and if changes are necessary, USAEEC will make recommended changes. The final verification score is established with the MOS monitor in conjunction with representatives of USAEEC.

If an individual fails to verify his MOS one year he can:
a. be retained in the MOS;
b. reclassified or;
c. have his skill level reduced.

A second consecutive annual failure to verify requires:
a. reclassification if failure is in the same Primary MOS or;
b. if in a different PMOS, the soldier may be retained until the next year and be tested again. A third consecutive annual failure regardless of Primary MOS involved in the three testings, results in the soldier being referred to boards convened under provisions of AR 635-209.

Nonverification of an MOS has distinct, and possible permanent repercussions on a man's Army career.

The next step or cutoff used in our system is an ES of 100. A linear transformation of the mean raw composite score results in an evaluation score of 100. The projected implementation date for this cutoff point will be February 1967. At this time, the upper 50% of the enlisted men in a given MOS, skill level, and pay grade will be eligible for promotion. This promotion will be termed a waiver promotion. This is an additional tool being implemented which will give the Army a management perogative; that of promoting an enlisted man, at his commander's discretion, who has scored in the upper 1/2 of his peer group.

An additional consideration made of all soldiers who receive promotions is where they will stand in relation to their new peer group. A recent study of waiver promotions indicated that nearly 90% of the waiver promoted soldiers will fall at or below the 50th percentile in the next higher pay grade. It is not certain whether such soldiers can compete with, or attain equivalent technical proficiency, as those already in the grade into which they have been promoted. However, if the commander feels a man is qualified despite his scores on the evaluation instruments he will be able to recommend promotion.

The final cutoff in this system is the minimum qualification promotion score. In order to be eligible for promotion,

a soldier must attain an evaluation score of at least 110. A soldier achieving this score is considered highly proficient in his MOS. The cutoff is made to include only the upper 1/3 of the personnel within any given MOS, skill level, and pay grade. Of course there is no immediate promotion unless the opening exists, but the promotability of the individual is assured for the entire year. The promotion, in the final analysis, is done by the commander from among those who have qualified at this level of proficiency.

Again, a second consideration of promotable enlisted men is where they stand in relation to the next higher pay grade. Generally, the upper 1/3 of a pay grade is at least comparable to the upper 1/2 of the next higher pay grade. Strong consideration is given to cases which fall below the mean of the next higher grade. This additional requirement may possibly eliminate a few individuals in any given MOS, but generally a score of 110 places the soldier well above the mean of the next higher pay grade.

EVALUATION DATA REPORT

Possibly the next point to bring out is our method of transmitting MOS evaluation information to the individual soldier in order to make it possible for him to better understand his strengths and weaknesses as indicated by the MOS evaluation test. We make no attempt to give each soldier a complete breakout as to the exact items he missed, or the number of items answered correctly. The basis for providing results on the MOS evaluation test is by highlighting subportions of the evaluation test called Broad Subject-Matter Areas. For instance, a Track Vehicle Mechanic is concerned with BSMA's such as: I - Army Maintenance Records, Reports, and Technical Publication; II - Tools and Equipment; III - Operating Principles of Internal Combustion Engines and Auxiliary Equipment; IV - Prevention and Organizational Maintenance of Internal Combustion Engines and Vehicle Electrical Systems; V - Prevention and Organizational Maintenance on Chassis and Power Train Assemblies; VI - Operation and Maintenance of Recovery Vehicles; and VII - Prevention Maintenance on Automotive Electrical Systems. The evaluation data report (EDR) is subdivided or profiled on each subject matter area of the evaluation test according to five distinct categories: Very High (VH), High (H), Typical (T), Low (L), and Very Low (VL). The profiling system was designed to include approximately 20% of the EM taking the test in each categorical area. The categories give each soldier his relative standing within the group being rated. In essence, the EDR is a soldier's "report card."

EVALUATION TEST PROFILE SUMMARY REPORT

Evaluation test profile summary report possesses many of the same characteristics as the individual evaluation data report.

The idea of making summary profile comparisons by subject matter areas can now be made between different levels of command groups to aid in determining the various strengths and weaknesses of each. This comparison is accomplished by determining the percentage of enlisted men in the selected group which score significantly High, Typical, or significantly Low in each subject matter area, and thus comparing these summary percentages to the expected standards.

The profiling for this evaluation test profile summary report utilizes the same methodology except the two extreme categories on each side of the scale have been consolidated. Reducing the number of categories enables one to more readily note significant trends. The individual EDR you will remember had five distinct categories with each accounting for approximately 20% of the population in each category. The consolidated form will consider the Very High and High categories as one category accounting for 40% of the population, Typical will still remain 20%, and the Low and Very Low categories as one category accounting for the other 40%.

In interpreting the results, one should consider a pattern of trends which may be indicative of very good or very poor performance. Hopefully some idea of the training specialization needed to correct significantly low subject matter areas can be attained and the proper steps taken to correct the situation.

SUMMARY

In summary, the concept of a multiple cutoff system has been incorporated into the US Army's MOS Evaluation System. Each of the cutoff points plays a significant role toward the accomplishment of good, sound management decisions on the part of commanders and trainers. Through proper application of the Enlisted Evaluation System, all enlisted personnel can be assured of an objective evaluation, against Army-wide standards, of their qualifications. They can be assured of equitable treatment in the development of their career program. Each new development in the system should prove more conclusively than ever that each soldier's chance of advancement is limited only by his individual worth.

ARMY EXPERIENCE WITH FIELD SCORING OF PERFORMANCE TESTS

Charles D. Roberts
US Army Enlisted Evaluation Center

1. This paper contains a brief discussion of the Army's experience with field scoring of MOS performance tests in four areas: Typing, dictation (stenography), court reporting, and music.

2. Typing Tests.

a. The typing tests consist of short samples of military correspondence such as letters or disposition forms.

b. These tests are scored by test control officers and their assistants. They follow detailed instructions regarding scoring procedures which are published in the administration and scoring manual.

c. The examinees are required to type the material in the test line for line. This requirement is designed to simplify scoring.

d. Several spot checks indicate that scoring is generally reasonably accurate.

e. The clarity and completeness of the administration and scoring manual is crucial to the success of these performance tests.

3. Dictation (Stenography).

a. The dictation (stenography) tests consist of short samples of military correspondence which have been recorded at the speed appropriate to the MOS on magnetic tape. The examinee records the material played by the tape recorder and then transcribes his notes with a typewriter. The use of any system by the examinee except the stenomask is authorized.

b. These transcriptions are scored by test control officers and their assistants. They follow detailed instructions regarding scoring procedures which are published in the administration and scoring manual.

c. Only certain scoring elements (which are underlined in the manual) are scored. The scoring elements consist of phrases and short sentences.

d. Several spot checks indicate that scoring is generally reasonably accurate.

e. The clarity and completeness of the administration and scoring manual is crucial to the success of these performance tests.

4. Court Reporting.

a. The court reporter test consists of excerpts from an Army courts-martial which have been recorded at the speed appropriate to the MOS on magnetic tape. The examinee records the material played by the tape recorder and then transcribes this material with a typewriter. The examinee is permitted to use any system he desires. Most employ the stenomask. Direct recording is not authorized.

b. This test is scored by highly qualified personnel at The Judge Advocate General's School. Generally all of the Scoring is performed by one person. This, of course, makes for consistent interpretation of the scoring rules. These procedures are published in a scoring manual. A thorough check on the scoring of this test has been conducted by the USAEEC. This study revealed that the scoring was highly accurate.

c. Only certain parts of the case transcribed by the examinee are scored. Analysis has revealed that the score achieved on this sample is very highly correlated with the score achieved on the total transcription.

d. Here again, the clarity and comprehensiveness of the administration and scoring manuals are crucial to the success of this performance test. A report of test conditions from each test administrator enables us to get insight regarding the quality of the administration manual.

5. Music.

a. The bandsman performance test consists of two parts -- an advance preparation test and a sight reading test.

(1) The advance preparation test contains music which is somewhat challenging to the examinee. This test is distributed to the examinee at least thirty days prior to the day on which he is tested.

(2) The sight reading test contains music which is less difficult than the advance preparation test. The examinee is only given a few minutes prior to being tested to examine the music.

(3) The examinee records both tests on magnetic tape.

b. The examinee's recordings are scored by a Bands-

man Evaluation Board. Eight boards have been organized to score these tape recordings. Each board consists of three qualified band officers. Each recording is scored independently by the board members. A scoring work sheet is employed by each board member to assist him in achieving uniformity, objectivity, and exactitude. This scoring worksheet is designed for ease and facility of use. It is designed to enable the board members to concentrate on the music.

c. Comparison of boards and of members of boards with each other indicates that the scoring is generally reasonably uniform. For example, a study of scoring of a 25 member band by a board produced the following Spearman rho correlations.

(1) Member A vs Member B on AP Test .942

(2) Member A vs Member C on AP Test .882

(3) Member B vs Member C on AP Test .945

(4) Member A vs Member B on SR Test .956

(5) Member A vs Member C on SR Test .965

(6) Member B vs Member C on SR Test .983

d. The clarity and completeness of the administration and scoring manuals is crucial to the success of these performance tests. Also, the design of the scoring work sheet is most important.

6. Conclusion.

a. Field scoring has been generally satisfactory.

b. The most satisfactory scoring is achieved in the professional type MOS such as court reporter and bandsman.

BLANK PAGE

AN APPROACH TO THE MEASUREMENT OF SUPERVISORY SKILLS

William W. Wance
U. S. Army Enlisted Evaluation Center, Fort Benjamin Harrison
Indianapolis, Indiana

INTRODUCTION:

After giving the measurement of supervisory attitudes and skills considerable consideration, I began asking myself this question, "Why can't this area of human behavior be less complex and more easily evaluated"? Then I remembered how easily a plant section chief had handled a problem of replacing one of his foremen who had died. A brash young chap, who didn't have much chance of getting the promotion, asked the section chief if he could take "poor John's" place. To which the section chief replied, "It's alright with me if you can arrange it with the undertaker," that, of course, is oversimplification of supervision.

The comments I intend to make today very likely will contain little astounding or new material; however, I believe a look at the systematic approach we in the Enlisted Evaluation Center used in trying to crystallize some of our thinking in this area will prove to be a beneficial review for all of you.

BODY OF REPORT:

A critical look at our approach to the measurement of supervisory skills revealed that the following three major points were bothering us:

1. The test outline subject-matter descriptions.
2. The number and type of supervisory items we possessed.
3. The type and adequacy of available study references.

My presentation will be concerned mainly with what we discovered while taking a critical look at these three areas and what we are doing in an attempt to improve our measurement in the area of supervision.

First, in studying the test outline subject-matter description, we performed, basically, what might be called an "armchair" content validity study - although we did study a small amount of data from test validation and factorial analysis studies which are now becoming available from our Evaluation and Analysis Branch. The "armchair" term is appropriate since we did not perform any direct job analyses but studied pertinent job information and supervisory studies

already available. Now, I would like to describe briefly the job content information included in our investigation.

The Army enlisted job descriptions identify the skill level "4" NCO's generally as leaders whose jobs involve relatively detailed knowledge of the tasks performed at all subordinate apprentice and journeyman levels in order to coordinate and give direction to work performed. Also, the skill level "5" NCO's are identified generally as supervisors whose jobs involve a broad, general knowledge of the tasks performed at all subordinate levels in order to coordinate and give direction to work activities.

Next, we took a look at the Army manual on "Military Leadership" (FM 22-100). It gives the following description of supervision:

> "A good leader will undoubtedly find that he will spend much more time in the supervision of action which is underway than he will in the formulation and issuance of actions and orders. To supervise properly, he must develop sufficient proficiency in the area which he is checking to be able to recognize whether the work is proceeding satisfactorily. He must accomplish supervision in such a manner that the progress of work in support of actions and orders is checked without unduly harassing subordinates."

A more specific look at certain MOS descriptions revealed the following key words and phrases as evidence of what is expected of the superivisory NCO:

For a Radar Maintenance Supervisor - Inspector

Must be able to plan and instruct on-the-job training***
Must be able to organize and supervise***
Must be able to effect liaison***
Must be able to coordinate***

For a Nike Maintenance Chief

Must be able to organize and supervise***
Must be able to coordinate***
Must be able to conduct on-the-job training***
Must be able to apply sound management principles, coordinate***, and maintain a high level of unit cooperation***

For a Field Artillery Senior Sergeant:

Must be able to plan, coordinate, and supervise***
Must be able to properly evaluate***
Must be able to evaluate and relate***
Must be able to possess a high degree of leadership ability***, skill in leading men, and directing group activities***

These extracted key words and phrases help to give the general tenor of what the Army expects from its supervisory NCO's.

Next, we looked at thirty-three "Supervisory" test questions with statistically significant validity indices which were identified in six validation studies of MOS evaluation tests. The following three test item stems are samples of the thirty-three studied:

(Show Viewgraph #1)

1. A subordinate questions the wisdom of your decision. What action should you take?
2. Your unit has a low court-martial rate. This situation is probably a result of which factor?
3. You are chief NCO of a maintenance section whose interests conflict at the moment with those of another section. You and your counterpart in the other section decide to talk over the conflict. What can BEST be accomplished by such a meeting?

The thirty-three validation study test items appeared to be related to an aspect of supervision which could be titled "inter-personal relations".

Following the look at the validation study test items we looked at six "Supervisory" items with high factor loadings which were identified in a factor analysis study of a supervisory skill level in the Medical Specialist MOS. The following three test item stems are presented as samples:

(Show Viewgraph #2)

1. During on-the-job training programs, the demonstration of an operation to beginners should be followed as soon as possible by --

2. As an NCO you make a special effort to assist all subordinates in the performance of their duties. You feel it is your duty to oversee their work constantly and closely. Furthermore, you are convinced that you can achieve best results if you personally decide how all work is to

be done. What effect is this behavior MOST likely to have on the operation of your section?

3. When making a decision, you should give the LEAST consideration to which of the following?

As in the case of the validation study test items the factor study items appeared to be related to the "inter-personal relations" aspect of supervision.

After having looked at some of the major Army MOS sources related to supervisory job information, we decided to look at a few civilian industry sources of such information; especially factor analysis results if they could be found. One interesting article was found in the journal of "Personnel Psychology". (Vol. 9, No. 4, Winter 1956) this article was titled: "Factor Analysis of Rated Supervisory Behavior." A pertinent paragraph from that article states as follows:

> "One of the major problems in the selection, evaluation, and training of supervisors is defining the skills, attitudes, and abilities required for effective job performance. Certain abilities such as technical knowledge, scheduling, procedures, etc. may be readily apparent, while many of the skills and attitudes involved in the inter-personal relations with employees and management personnel may be difficult to determine."

Fifteen factors emerged from the analysis. They were:

(Show Viewgraph #3)

1. Open-Mindedness
2. A General Bias Factor
3. Personal Compliance
*4. Job Knowledge
*5. Direction of Employee Performance
*6. Rewarding Performance
7. Company Loyalty
8. Cheerfulness
9. Acceptance of Responsibility
10. Group Spirit
11. Approachability
12. Personal Drive
13. Impartiality
14. Poise
15. Consideration

The job knowledge aspect of supervision is the one usually included in paper-and-pencil tests of supervisory per-

formance. The inter-personal relations aspect of supervision is naturally more difficult to measure. Factors numbered 5 and 6 are the only other ones tangible enough in this area of inter-personal relations to lend themselves readily to paper-and-pencil testing.

More insight into the skills, attitudes, and abilities required for effective supervisory job performance was revealed by the National Management Association after a study of 86 companies. The following is a list of six supervisory pitfalls revealed by the study:

(Show Viewgraph #4)

1. Poor Personal Relations With Workers or With Other Management People.
2. Individual Shortcomings, Such as Lack of Initiative, Emotional Instability.
3. Lack of Understanding of the Management Point of View.
4. Unwillingness to Spend the Necessary Time and Effort to Improve.
5. Lack of Skill in Planning and Organizing Work.
6. Inability to Adjust to New and Changing Conditions.

It is interesting to note that the personal relations aspect of supervision heads this list of pitfalls.

After studying the background information we have just reviewed, three major points became evident: (1) supervisory skills, attitudes, and abilities may be broadly categorized into supervisory technical knowledge and the skills and attitudes involved in inter-personal relations, (2) supervisory NCO's do get involved and are expected to get involved in the inter-personal relations aspect of supervision, and (3) under most conditions of supervisory NCO performance, information presently available indicates that a good supervisor will spend more time in promoting individual and group performance than in informing subordinates concerning what is to be done or in following up on performance.

Since the technical knowledge aspect of supervision is already accounted for in much of our paper-and-pencil testing, we decided to make an attempt to isolate the "inter-personal relations" aspect of supervision in a separate test outline subject-matter area. Previous versions of our test outline subject-matter areas covering supervision often had been a blending of the technical

knowledge and the "inter-personal relations" aspects of supervision. For the area of Supervision (Inter-Personal Relations), we decided upon the following three major subheadings:

(Show Viewgraph #5)

1. Informing Subordinates Concerning What Needs to be Done.
2. Promoting Individual and Group Performance.
3. Following up on Performance.

As an interesting comparison, we can look at the three main elements of The Management Cycle used by the American Management Association in an Intensive Supervisory Development Course. The elements are:

(Show Viewgraph #6)

1. Planning
2. Action
3. Control

An obvious result of the comparison is the high degree similarity in the two lists of supervisory elements.

Our complete test outline subject-matter description for "Supervision (Inter-Personal Relations) contains additional detailed and descriptive subject-matter under each of the three major subheadings we just reviewed; however, the coverage of these details is not important to this presentation.

A last point to be made about the test outline subject-matter concerns the matter of "item weighting" for the three major subheadings, i.e. assigning the number of items to be written for each area. Based upon the background information we studied, it appeared appropriate, on an "a priori" basis, to assign twice as large an item weight to the subheading "Promoting Individual and Group Performance" as to each of the other two subheadings. The weighting structure will be checked further as empirical data become available. With the weighting factor determined we have covered the basic handling of the test outline subject-matter by the US Army Enlisted Evaluation Center.

Next I would like to make a few comments about selecting and developing a "bank" of 200 supervisory test items appropriate for the revised test outline subject-matter. I think all of us are aware of the difficulties

encountered when we attempt to construct items to evaluate supervisory abilities. We at the USAEEC were not too happy with the content of some of the items we had and with the number of items from which we could select in order to have alternate sets of supervisory items. In looking for items to fit our revised test outline subject-matter, we found the following sources and selected, revised, or developed items appropriate to the test outline requiremnts:

(Show Viewgraph #7)

1. USAEEC Supervisory Items Presently Being Used,
2. Army Adjutant General's Management Course Test Questions.
3. MOS Validation Studies.
4. A Factor Analysis Study.
5. New Items Constructed by USAEEC Test Specialist.
6. Air Force Supervisory Practices Test.
7. Air Force Squadron Officer Course (NRI) Test Questions.
8. Navy E-8 Test Which Included Some Supervisory Questions.

I would like to discuss briefly a new and novel type of test item included in the "bank". It is a situation type item designed to have the examinee "project" his supervisory attitude or skill into the answering of the item without being aware tnat he is relating information about his supervisory behavior. The following is a sample of such an item:

(Show Viewgraph #8)

SITUATION:

Sgt Yankee is an instructor of a special course on "Leadership for NCO's." Sgt Mike, a member of the class, monopolizes the class discussion time and prevents adequate participation of others in the class. Sgt Yankee has tried various methods to reduce Sgt Mike's talkativeness in the class sessions but has not been very successful. He decides to talk personally with Sgt Mike about the problem he is creating. The discussion begins with the following comments.

Sgt Yankee: "Sgt Mike, I would like to discuss the type and amount of classroom participation which we have had in our Leadership class. I think you are aware that you do a great deal of the talking which occurs during the class session. I appreciate your willingness to participate, but ---."

Sgt Mike interupts Sgt Yankee: "Oh, sure. I'm aware of it. I'm very much interested in the course and the other fellows don't seem to have much to say, so, I make my contributions."

Sgt. Yankee:

(Show viewgraph #9)

Assume that you are Sgt Yankee with the information and conditions stated above. What would be the BEST response to make to Sgt Mike at the point where the interview stopped?

A. "Well, getting participation from all members of the class is part of my job as instructor."
B. "What do you think we could gain by giving the others a little more time to talk?"
C. "Don't you think the others may desire to talk but are usually cut off by you?"
D. "Individuals differ in their aggressiveness in the classroom situation. Some need more encouragement to participate than others." The "situation" part of this item was designed such that the supervisor was not given all the facts or information necessary to reach an immediate decision or solution to the problem. He had to recognize this lack of information and also sense what would be the next best response to give to draw more information from the interviewee. Based on these two conditions, the best response to the sample item is "What do you think we could gain by giving the others a little more time to talk?" To those who are acquainted with "non-directive" counseling or interviewing techniques you will note a great amount of similarity.

One might suspect that examinees would "catch on" to the built-in projective technique used if such items were presented in a series in a test; however, a group of six editors and one female supervisor of typing personnel answered ten such items in a series and none detected the similar "projective technique" which was inherent in each item. In an ordinary test such items would be mixed with the more conventional type of items and, therefore, would be even less subject to having the pattern detected. None of these items has been tried out empirically but we hope to obtain positive results along with trying something different.

It is my sincere belief that any major "break throughs"

in the structure of multiple choice test items to evaluate personality characteristics or the attitudes and skills related to "inter-personal relations" of supervision will come through the use of some sort of "projective" type test item such as the one just described. The surface appears to have been only scratched in the area of test item structuring.

Now, for a few comments about the last of the three major points I set out to discuss, namely, "The type and adequacy of study references available to cover the area of supervision for the NCO level of performance." We soon became aware that we were at a disadvantage in this area and I was reminded of the story about the golfers. A foursome of golfers barely had begun their game when they noticed a shapely, nude girl come dashing across the greens nearby followed by four men dressed in white coats. One of the golfers managed to get the fourth man stopped to inquire about the unusual performance. The man in white answered that there was a mental institution a short distance away in which the girl was a patient. Periodically, she got a desire to undress herself and run away. The golfer replied that the explanation sounded very logical but he was still curious about the third man in white who was carrying a bucket of sand. "Oh, that's very simple replied the man in white. "That's his handicap he caught the girl the last time."

We are handicapped by lack of adequate study reference material for the "inter-personal relations" aspect of supervision appropriate to the NCO level of supervisory performance. The major Army reference is a field manual (FM 22-100) titled "Military Leadership." This manual is not an especially good one since the material is slanted toward the officer or "commander" level of performance. Also, a large portion of the material refers to the less tangible factors of supervision as far as paper-and-pencil testing is concerned. Several non-resident instruction courses contain material either directly or indirectly related to inter-personal relations appropriate to NCO performance; however, this study material is not available through normal DA publication distribution channels and it gets listed in the test aid study reference listing only as suggested study material. A disadvantage of much of the pertinent non-resident instruction material is that of being written at a level above the normal operational level for a supervisory NCO. For example, the term "management" can have many interpretations or meanings depending upon the background of the individuals using the term or the particular environment or setting for which the term is serving as a "referent." Normally, enlisted personnel think of the term

"management" as referring to the officer level and civilians think of it as referring to the executive level.

A formal suggestion has been submitted through channels in an attempt to stimulate the development of a more appropiate Army manual for the subject-matter area we have been discussing. In the meantime we would like to make an appeal to each of the other Armed Services represented here today to inform us of any appropriate reference material their service has available for the NCO level of supervisory performance. In fact, we are willing to do some "horse trading" by exchanging a copy of our compiled "bank" of supervisory items for copies of pertinent references related to the area we have been discussing.

In this presentation I have attempted to give you some of the recent USAEEC thinking and action taken in the following three major areas:

1. The test outline subject-matter descriptions.
2. The developing and compiling of a "bank" of supervisory test items.
3. The identification and tentative solution of the study reference problem.

I realize that in a highly complex area such as "supervision" we vary greatly in our viewpoints and beliefs. However, I immediately think of the frequently made comment that if two individuals or two parties are working and thinking in the same subject-matter or problem area, and if both parties are in complete agreement there very likely is no need for one or the other of the parties concerned. I prefer to believe that each of us is needed in our respective evaluation efforts.

ABSTRACT OF PRESENTATION

Dr. Charles A. Pounian
Personnel Officer, Chicago Civil
Service Commission

The City of Chicago, Civil Service Commission is the central personnel agency for the City of Chicago government. It also has jurisdiction over the non-teaching employees of the Board of Education and the Chicago City Colleges. This involves a total of about 48,000 employees.

The City of Chicago has been faced with a difficult recruitment problem for the position of Patrolman for the past several years. A particular aspect of this problem has been the inability to obtain the services of Spanish speaking officers in sufficient numbers to meet the needs of the Police Department. This situation was dramatized in May of 1966 by the disturbances in the areas on the northwest side of Chicago predominantly populated by Spanish speaking citizens. It was felt both by the Spanish community and the Police Department that a greater number of Spanish speaking police officers could have either avoided the problem or reduced it in scope once it had begun.

The Civil Service Commission, working closely with the Police Department, established a liaison with the Puerto Rican community in a number of ways in order to solicit their assistance in interesting young Puerto Ricans a career on the police force. In April 1966, prior to the disturbances on the northwest side, the Commission had already taken several actions which made it easier to attack the problem. The Commission announced that the Patrolman examination would be given on a weekly basis for an indefinite period. The examination had been given only every three or four months prior to this.

At this time the physical standards were reviewed and in some cases eased. A minimum height requirement was lowered from 5'8" to 5'7". This change in particular allowed a much greater number of Puerto Ricans to qualify for the examination. This minimum height is the lowest for the position of Patrolman in any major city in the United States.

The Commission also began work on a pamphlet which will provide information to prospective Patrolman candidates on specific item content of the test. The booklet will give three examples of each of the six different kinds of items on the test, explain them and then give the correct answer together with an explanation as to why it was correct and the other answers were incorrect. The purpose of developing

this material was to acquaint persons who were not familiar with multiple choice material with the type of material they would find in the test. Patrolman candidates raised in large metropolitan areas in the United States such as Chicago are very familiar with multiple-choice testing. However, American citizens from Puerto Rico are not accustomed to multiple-choice testing. Because of this they operate at a disadvantage in the testing situation which we feel will be reduced through careful review of our pamphlet. We are also translating this pamphlet into Spanish (with the exception of the test items) in order to give as great an understanding to Spanish speaking candidates as possible. While the results of our efforts have not been dramatic there has been a steady increase in Spanish speaking Patrolman candidates and therefore a greater number of eligibles.

The Chicago Civil Service Commission has traditionally used an education and experience rating on original entrance examinations. In recent years we have re-evaluated our testing techniques to make sure we were not eliminating people for reasons other than the ability to do the job. As a result of this re-evaluation we have dropped from consideration education and experience in a variety of original entrance examinations including Patrolman, Firefighter, Typist and Junior Stenographer. In the case of Patrolman and Firefighter we are attracting persons who are capable of benefiting from an intensive training program. Our mental abilities examination in these areas has proved to be an effective predictor of success in the training academy for these positions. In a recent study of Patrolman candidates, 65% of persons with at least some college credit passed the Patrolman test and were placed on eligible lists. Approximately 35% of those with high school but no college passed the test, and 20% with less than high school graduation passed the test. Based on previous standards these high school drop outs would have been rejected; however, they performed better on our examination than many persons with some college training. In the case of Patrolman we have used this approach for approximately five years to the satisfaction of the Chicago Police Department. We have actually raised our mental ability standards, but because we have broadened our potential population, we are able to attract more qualified candidates than before.

Our policy of eliminating the evaluation of education on Patrolman examinations has been extremely helpful in our attempts to attract Spanish speaking police officers. Many Puerto Ricans who applied for the examination are high school drop outs who would, under the old standards, be eliminated. In fact, a private consulting firm which undertakes an independent evaluation of police applicants with

primary emphasis on personality evaluation gave its highest recommendation to a Puerto Rican who was a high school drop out. This candidate, who performed well on the examination, would not have been considered under the old standards.

In the case of clerical examinations we are able to determine clerical ability, typing skill, and/or ability to take shorthand, as required by the job in question. The introduction of an educational factor is really unnecessary. As a practical matter the great percentage of successful candidates for these jobs are high school graduates. The rigorous testing procedure eliminates the persons who are not capable of performing the duties.

We have a responsibility to review our standards and techniques to make sure that we do not eliminate persons in the selection process who are capable of performing a job.

In many cases the requirement of high school graduation is artificial. Other evaluation techniques found do a much better job of discrimination between the qualified and unqualified candidates.

SYSTEM ANALYSIS APPLIED TO PERSONNEL RESEARCH

J. M. Keenan
Stanford Research Institute

The application of systems analysis to personnel research should lead to optimization of job satisfaction for the individual, on the one hand, and increased efficiency for the organization on the other. Since a large percentage of system error has been shown to be due primarily to the human factor, behavioral scientists should be called on to study the problem of human variance. From a cost/benefit viewpoint, concentration on the human component as a cause of variance in systems behavior produces maximum economic returns. An analysis of the relationship between individual and "culture" profiles should facilitate the process of matching men with jobs in an organization. A close look at current job selection procedures (testing techniques) may also indicate a need to restructure certain aspects of the organization.

Systems Analysis--the Long Run View

One of the better concepts of systems analysis is:

> . . . operations research is designed to optimize around resources and constraints at the moment; systems analysis attempts to take a larger, long run view by integrating elements from outside the local system with the constraints placed on the existing system. It then examines cost/effectiveness methods of changing these constraints.*

In general, the elements of systems analysis are: identification of an objective, identification of alternatives, examination of costs, a model of the system, and a specified criterion function. The analysis is then iterated. Models enable us to handle multivariate optimization or, in cybernetic terms, to cope with exceedingly complex probabilistic systems. The current state of the art is such that it is now possible to examine organization systems with the tools of systems analysis.

System Variance due to Human Error

The ultimate objective in using systems analysis in personnel research is the matching of men and jobs to the optimal joint satisfaction of individual and organization. As the physical and engineering sciences increase the precision

* Jack Stockfisch, Special Assistant to the Chief of Staff of the Army.

and accuracy of their predictions, it becomes evident that the major source of variance in systems behavior is due to the human component. Scientists are interested in developing better techniques for handling the indistinct problems inherent in all areas where human behavior is a major element. Descriptive and predictive models of organization behavior have been developed for this purpose.

The continuing dissatisfaction with the behavior of complex organization systems leads to appraisal of human performance error or assigning of blame. If the world was binary, comprising people who were either "winners" or "losers," there would be a simple, probabilistic model for the personnel assignment problem. Unfortunately we are required to make judgments that imply an interval scale of measurement, based on sorting devices that tend to be nominal at best. The basic argument presented here is that the major function of appraisal is descriptive and should be concerned with attacking the problem of variance with maximum efficiency.

The greatest amount of variance probably can be accounted for by values, goals, and personality traits. The human tendency to attack simple, but perhaps trivial, problems could lead to extensive arguments on research strategy and inductive and deductive logic, and, in fact, these arguments may have been very relevant ten years ago. Today information technology is sufficient to handle large scale, complex design. For example, Saul B. Sells (1964), at Texas Christian University, is working with a factor analytic structure of personality involving 600 variables. He presumes the model to be highly redundant but when problems of this scope can be handled by modern computing methods, "fishing" expeditions become interesting.

On medium-sized computers, fairly large scale stepwise multiple regression and multiple discriminant models can be handled. Considering the exceedingly rapid computation time, a thoughtful research design and multiple runs will allow the researcher to handle many more variables than the 50 to 100 that are routine on general purpose library programs.

The major problem is one of optimal matching of men and jobs--selection---based on more sophisticated systems than have been utilized previously. We are using rather gross, operationally defined criteria for personnel allocation. For example, the criteria for selection for officer training tend to be crude, with unknown relevance to long term performance (e.g., the high attrition rate). Also there are seldom upper level cutting scores on intellectual ability or other discriminations of this type.

Is ongoing appraisal the answer? Perhaps this question

begs the issue. The process of selection and matching may be the problem, rather than performance appraisal, which should only be employed for error correction and identification of growth. To reiterate, we can account for more variance by examining selection procedures rather than appraisal procedures. If selection is made properly on some sort of optimizing basis, appraisal will not be a particularly significant portion of the problem.

Cost/Benefit--The Human Element

There is a strong temptation to fund projects that can be precisely described, particularly where the expected results appear to be amenable to higher order measurement techniques. In many cases, studies of this type are inappropriate because they do not attack those elements that account for the major portion of variance. For example, the army has spent hundreds of thousands of dollars determining, almost to absurdity, the relative accuracy of various small arms; however, little has been spent on investigating why a very large percentage of the soldiers in Korea presumably either did not fire their weapons or fired into the air. If the principle of suppressive fire is relevant, then extreme accuracy is trivial as far as total effect is concerned, and if suppressive fire is not relevant, the major source of variance is the man.

This position may be challenged by the following analogue. There are several factors involved in the performance of a good athlete, such as conditioning, aptitude, and equipment. While two different athletes may not have equal ability, there is no reason why the equipment variable cannot be equalized. Based on this argument, fantastic sums of money are spent on examining alternatives in small arms weapons systems. This attack is relevant only after alternative uses for research resources are determined.

If this problem is viewed in terms of cost/benefit analysis, it becomes apparent that the maximum expected gains are to be found in studying the human element, which accounts for the greatest percentage of variance in systems performance. In fact, ultrahigh accuracy of the weapon will probably be a much less significant variable.

Given the state-of-the-art, behavioral studies are of greater relevance to man-machine systems performance than physical science studies. We should be interested in the greatest payoff in systems performance ("the most bang for a buck") as compared to the cost of research. Marginal efficiency implies a shifting of funds into improving behavioral research.

Quantitative Models of Human Behavior

History

Mathematical modeling of human behavior is attractive and has been employed for some time. Mathematical models are clean and yield precise predictions, although historically quite inaccurate. The Lewinian system (Lewin, 1936), using some principles of topology, is an attractive representation of personality and appears to explain a great deal about human behavior. Alec Bavelas (1950) provided an early theoretical model of group structure using what is now known as "graph theory." These pioneers have had a substantial influence on researchers in this country, and there is a substantial current effort to provide mathematical descriptions of human performance, both individually and in groups.

Ashby (1960, 1961) and Beer (1959) suggest that cybernetic models should be applicable to human behavior in organizations. They provide a basic quantitative framework for describing systems of this type (see Ashby's An Introduction to Cybernetics). They consider the probabilistic nature of systems, and suggest some methods of contending with at least risk, if not uncertainty.

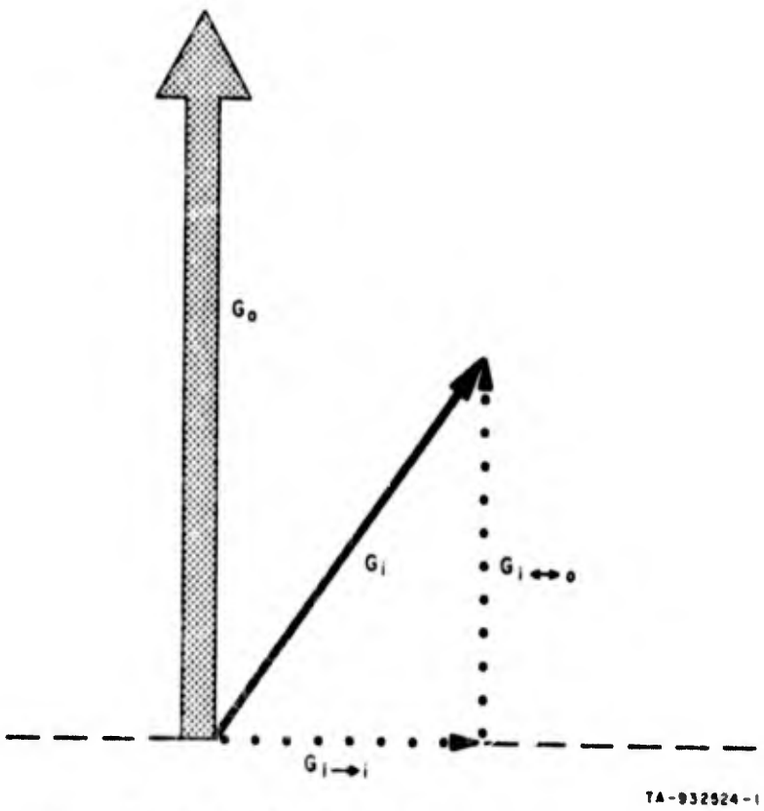
A Vector Model of System Performance

In Figure 1, the matching process between organization and individual is represenced in two dimensional space. The theory can be extended into more dimensions as appropriate. The primary vector, Go, represents the goals of the organization; the secondary vector, Gi, the goals of the individual. For those familiar with factor analysis, these can be viewed as a factor or cluster of elements. There are a number of choices in respect to appraisal, selection, retention, and retraining. When vector I is broken into its components, a vertical vector corresponding to the organization goals ($G_{i \to o}$) and a horizontal vector corresponding only to individual goals ($G_{I \to I}$) are found. The task of testing is to determine the relative spatial location of G_i; a partial task of training would be to reduce the angle between G_o and G_i. Efforts to motivate would have to be made to increase the length of G_i or $G_{i \to o}$.

Appraisal or testing, correctly used, is the most important element in a systems model for personnel utilization. Many quantitative techniques are sufficiently developed for use at present. Mathematical programming is becoming increasingly sophisticated and, assuming adequate inputs are available, there will be no problem in developing better models than those in existence. The suggestion is, therefore, that research be shifted toward appraisal, particularly in deriving better descriptions of human values and goals.

Figure 1

A VECTOR REPRESENTATION OF INDIVIDUAL AND ORGANIZATION GOALS



G_o = GOALS OF ORGANIZATION
G_i = GOALS OF INDIVIDUAL
$G_{i \leftrightarrow o}$ = PURE CONTRIBUTION OF INDIVIDUAL GOALS TO ORGANIZATION GOALS
$G_{i \rightarrow i}$ = INDIVIDUAL GOALS INDEPENDENT OF ORGANIZATION GOALS.

Personality

Values and the Task

If organization design is a problem that will yield to optimization tactics, we will need operationally defined variables for the model. Only two major variable categories will be discussed: individual personality and organization personality and the contrast between them. Operational descriptions are needed of (1) the relevant dimensions of the individual and (2) the organization system in which he must operate; the current level of technology should be able to handle these problems. Although rather a subtle distinction, the significant consideration is how well the man is matched (in values) to the task. If the job is essential, we have a system that needs to be optimized across the labor resource rather than across the task structure. The major concern is a probabilistic estimate of how well a man of a particular type does a particular job, on the average.

Interdependent Trait Sets

One way to classify personality is to group traits into interdependent sets, such as congruent, complementary, or competitive. The congruent set is measured by a high, positive correlation between the scores of two individuals where the trait is mutually fulfilling, e.g., affiliation, loyalty to friends, and participation in friendly groups. The complementary need category is a mutually interacting force between two negatively correlated needs. Examples on the Edwards Personal Preference Schedule (EPPS) are Aggression and Abasement. Agression is an attacking behavior and Abasement is a negative syndrome characterized by guilt, timidity, and inferiority. Competitive characteristics are traits that would conflict, e.g., Dominance-Dominance, or Dominance-Aggression. Schultz (1958) demonstrated the feasibility of predicting behavior based on the interdependence between two individuals. Keenan and Hoverland, (1965) demonstrated the relevance of the interaction effect of personality variables at the U. S. Naval Postgraduate School.

The task is to identify the positive aspects in the environment, i.e., why officers stay, not why they quit. Perhaps the work should be based on a modified critical incident technique, identification of particular points at which the individual most liked or disliked his environment.

The above system sounds effective; however, it contains a trap, which must be avoided. As far as the organization and individual profiles are consistent*, we could have a highly

*Robert Anthony, ASD (Controller), calls this "Goal Congruence"

predictive matching model (Figure 2a). However, under the current rate of change in technology, the existing profile should be adapted accordingly. The key to the problem may be organization adaptivity; therefore, provisions should be made to allow for the deliberate introduction of variance in personality types so that the organization structure can remain viable and adaptive to changes in the environment (Figure 2c).

Normal Distributions

We presume that some sort of normal distribution of individual variables exists in human personality. Within subsets of types of people, there seems to be somewhat less variance in the inherent types. My work has been with naval officers but Janowitz (1960) states that this would apply to all military officers. In our study of a selective subset of naval officers with a sample size of over 400, Prof. John Senger, Prof. H. Arthur Hoverland, and I have found that these officers have remarkably consistent personality profiles. Generally they will tend to test slightly high on dominance and aggression, but the striking element about the profiles is how average they are. The "form" of each profile is remarkably consistent from officer to officer, i.e., the variance around each point estimate is very low.

It is possible to discriminate exceedingly well between naval officers and, for example, students. A multiple discriminant analysis, done on the Edwards Personal Preference Schedule (EPPS) between officers of the U.S. Naval Postgraduate School and college juniors at Western Michigan University in Kalamazoo (Keenan, 1966), separated groups remarkably well. The only overlap of naval officers into the student group comprised two subsets--Waves and, interestingly, male naval officers who were undergraduates. (Sample sizes were 200 and 130.)

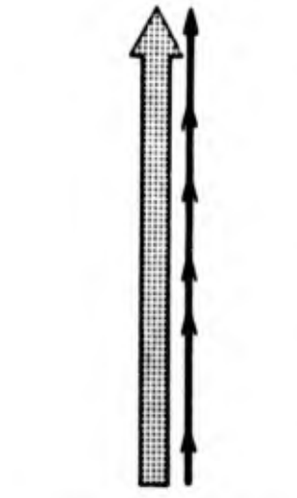
With the consistent profiles derived from the tests at the Postgraduate School, it may be feasible to identify what is now considered to be a "successful" junior officer.

Individual and "Culture" Profiles

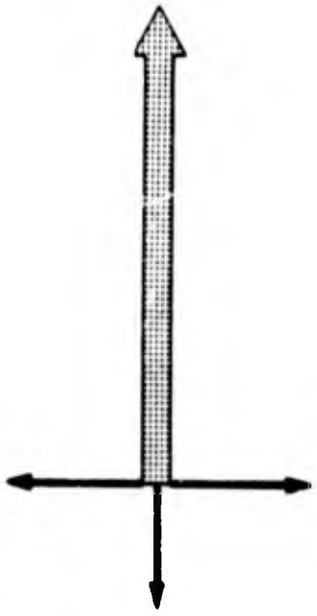
The upper curve shown in Figure 3 is from the Heterosexuality scale on the EPPS. Out of 115 subjects, tested in one year at the Postgraudate School, the lowest subject on the Heterosexuality scale was at the 85th percentile for the population (the highest score was made by a very popular girl). This scale on the Edwards test gives items with forced choices, such as: Would you rather kiss members of the opposite sex or read a book? It appears that any self-respecting naval officer knows the "right" answer to that question. We suggest that it is a matter of knowing the "right" answer to the question and that girls and books have nothing to do with it.

Figure 2

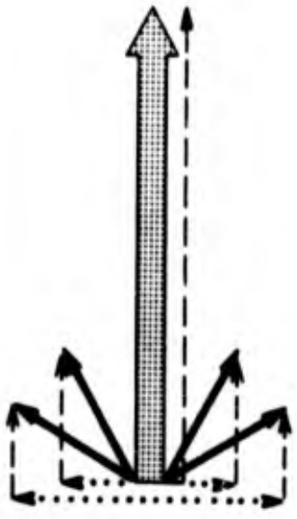
POSSIBLE VALUE CONFIGURATIONS IN ORGANIZATIONS



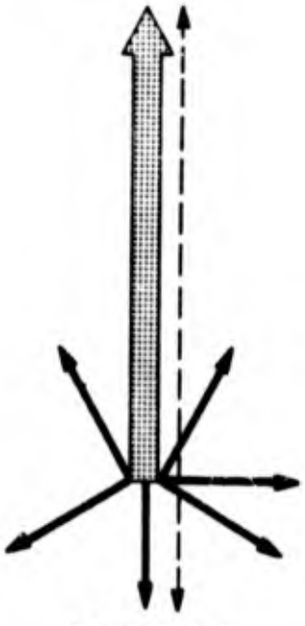
(a) GOAL CONGRUENCE



(b) DYSFUNCTIONAL



(c) ADAPTIVITY



(d) RESEARCH

TA-932824-2

On another test (Minnesota Multiphasic Personality Inventory) naval officers rated slightly high on femininity. The inference is that one of the tests measures culture while the other measures individual values. Considering the nature of the questions, we feel that the Edwards test is measuring the culture. For completely secure males, the question, "Would you rather kiss girls or read books?," is very difficult to answer; it depends on the girl and the book.

The problem now is the appropriate matching of individual to organization for maximum joint satisfaction, described in the economic literature as Pareto optimality.

Restructuring the Organization

The retention problem probably could be handled best by restructuring the organization in such a way that more or better people are initially put into the manpower channel. This would eliminate the necessity of attempting to stop up the channel. In other words, the retention problem is really a selection problem.

Success and Selection

Preferably, selection criteria should be related to success criteria. Typically, however, major selection criteria are based on success in jobs that may be irrelevant to the future job. Janowitz (1960, P. 11) adopts this as one of his major hypotheses related to attaining flag rank. He suggests:

> Prescribed careers performed with high competence lead to entrance into the professional elite, the highest point in the military hierarchy at which technical and routinized functions are performed. By contrast, entrance into the smaller group--the elite nucleus--where innovating perspectives, discretionary responsibility, and political skills are required, is assigned to persons with unconventional and adaptive careers.

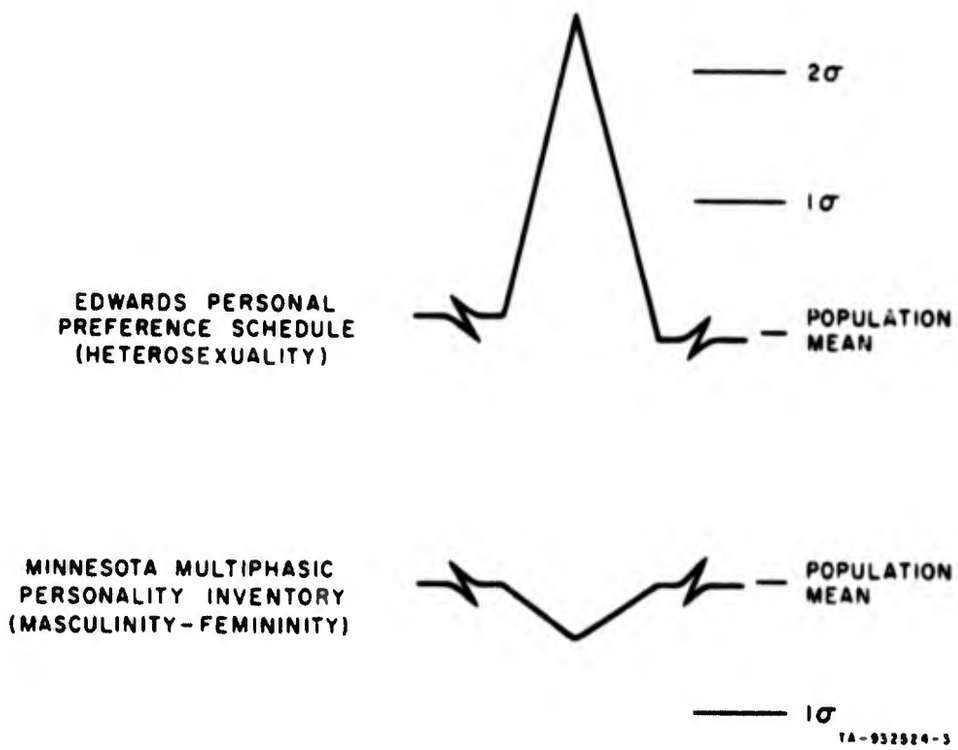
While at the U.S. Naval Postgraduate School, I suggested the following model (met with outrage by all-but-passed-over Commanders):

1. Those acts that suggest innovative executive ability are precisely those acts that decrease the probabilities of the man moving to the top levels of upper middle management (Captain), and

2. Those acts or behavior sets that enhance one's chances of moving to upper middle management are precisely the acts that suggest lack of executive talent.

Figure 3

REPRESENTATIVE PROFILES FOR SELECTED NAVAL OFFICERS
ON TWO PERSONALITY SCALES



BLANK PAGE

In other words, as you perform acts that increase your chances of becoming a Captain, you decrease your chances of becoming an Admiral (given that you make Captain).

Another great difficulty within the military structure is the cost of "accidents" to career patterns. For example, if a Naval aviator dumps a plane into the drink three or four times on takeoff, it seriously decreases his chance of moving to executive levels (even holding the attrition rate constant). This is an extreme example, but it indicates the need to question the quality of certain filtering systems and their efficiency in bringing to light and developing executive talent.

Means of Accomplishment

Political scientists (in particular, Thompson, 1961) have suggested three ways of getting things accomplished: power, authority, and influence. Janowitz suggests an increasing need for the use of influence within the military structure, especially since the military operates most of the time in a bureaucratic structure, rather than under crisis.

In many instances a military leader who has the authority to make a decision finds that he does not have the necessary technical expertise to actually make the decision without calling on his technical subordinates. Thus command positions are increasingly involved in the coordination of these technical specialists. The performance of specialists cannot be easily assessed, which gives them a definite "power" advantage.

The technician in our society can have a considerable amount of power simply by withholding his abilities. An electronics technician does not have to use the most efficient search procedure; a radar observer does not have to observe as quickly; a scientist does not have to think. A classic example occurred during World War II when an airframe manufacturer's design engineers were all drafted and placed back in the same design jobs, with corresponding pay cuts of about $400 a month. Thereafter, these men seemed to have difficulty creating new designs as quickly as before. Withholding technical expertise is a major source of power and an organization design in which authority depends on this power tends to concentrate this power even more.

Power may be used more frequently when diffused. The man who has a stratified (50% computer programming, 30% clerk, 20% other) job is likely to use his power to create a favorable shift in his time usage (perhaps to stay off K.P.), whereas the man who is practicing a profession (using the word loosely), is more inclined to use influence in lieu of power.

Stratification

Under the present job structuring system, a given man has quite a wide range of tasks to perform, as shown in Figure 4. These may range from high order intellectual tasks, such as computer operation, down to such things as kitchen duty. Looking at the system, man by man, quite a wide variety of skill levels are utilized by one man during his working month. From a systems analysis viewpoint, there is no category called "other duties." A man's job would be viewed in terms of the aggregate time expended during a year, regardless of how it was expended.

A more appropriate allocation between men and jobs might be to stratify, that is, to have all higher skills practiced by some and all lower level skills practiced by others. The implication is that job satisfaction is more important than any other motivating force. If a man is required to perform, or better, allowed to perform in accord with his personal values and goals, discipline and maintenance problems will be reduced.

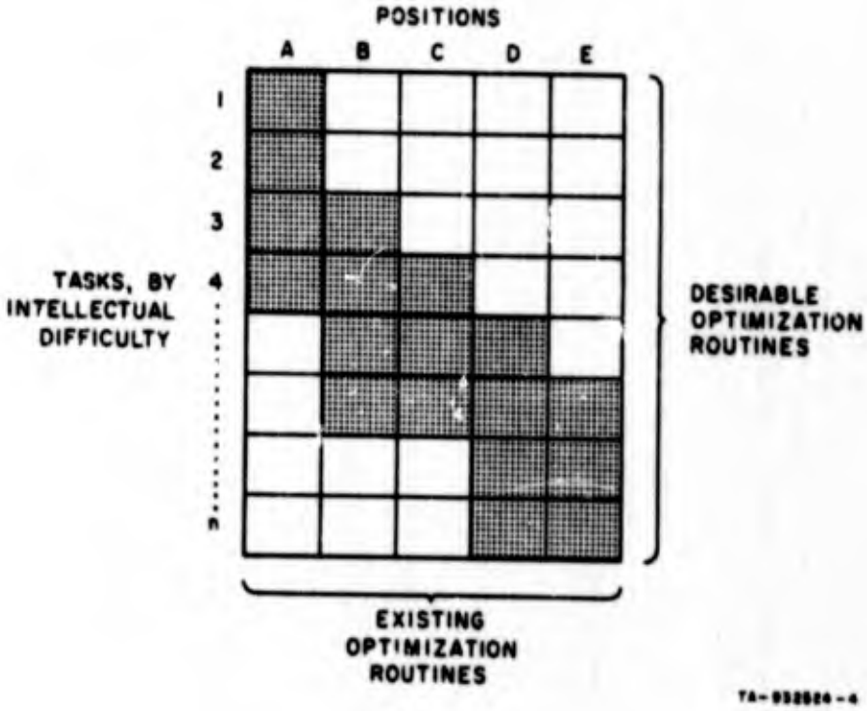
This theme would probably be viewed as prejudice by the parents of America, particularly those whose sons are potato peelers. However, this system leads to a much higher level of aggregate performance than is possible with the present system. This structuring allows for efficient performance of tasks with less manpower. Presuming the model of a man being allowed to work to his capacity, job satisfaction goes up, and there is a considerably better allocation mix, reducing the manpower requirements. Naturally, the possibility of restructuring the organization in this manner does raise some questions in respect to manipulation of the organization by people at the lower levels.

Concluding Remarks

Through skillful use of the information processing techniques developed by operations research people, multivariate statistical models, and the substantial information processing capabilities of modern computing systems, a technology mix can be developed that will lead to significant advances in the utilization of personnel resources. The major problems at this stage are probably conservatism and the lack of willingness to move to viable organization designs--problems of feasibility as contrasted with possibility. The military has tended to be very conservative and has focused on the problem of retention in an economic sense (pay raise). It may be the time to view research dollars in light of cost/effectiveness measures. Substantial strides should be made in organization design and the optimal matching of capabilities to organizations in the near future.

Figure 4

POSITION-TASK MATRIX



Systems analysis techniques can, and should, be adopted to reduce system variation through better utilization of the human resource. Assessment technology, coupled with quantitative techniques, should be developed to approach the problem of optimal organization designs.

A RAPID SOLUTION FOR MULTIPLE R AND BETA WEIGHTS, WITH CHECKS

John S. Brand
US Army Enlisted Evaluation Center, Fort Benjamin Harrison
Indianapolis, Indiana

Thorndike (1949) gives an abbreviated Doolittle solution for multiple R and beta weights (App. A, pp. 336-339). DuBois (1957) also gives an abbreviated solution (pp. 45-53). The purpose of these solutions is to reduce labor and rounding and transcription error by accumulating a number of multiplications in the calculating machine, recording only the final result. A disadvantage of the abbreviated method is that there are fewer residuals which can be used for special analyses. These solutions are essentially the same, except for minor variations. Thorndike uses three rows per block, while DuBois arranges the third row (betas) vertically in the space available on the left of the worksheet. Thorndike omits the last series of subtractions in Block 0 which gives $1 - R^2$. Both provide the usual checking procedure for each row of the operation. Thorndike places his final betas at the bottom of the criterion column, and DuBois places his final betas at the left of the worksheet. In both solutions, the order of calculations is somewhat difficult to follow.

Putting the two systems together, using the three row format, pointing out a principle for describing the more complex stage of the operation, and rearranging the back solution provides a rapid and simple method for solving for multiple R and the beta weights. The solution is simple enough that conventional checking methods for each row of the operation have been omitted. Checks are provided for the final solutions. If they do not check, it is not difficult to run through the entire solution again. In the first solution, reasonable care will nearly always provide accurate results.

The results given are exactly the same as those obtained from the more lengthy Doolittle (called "Do-much" by some) solution. Results will actually be more precise because they are less subject to rounding error. The method requires conventional calculating machines, and is good for up to about ten variables. Over ten the work becomes excessively tedious and more subject to error. The majority of multiple correlational analyses involves less than ten predictors. For more than ten variables, any manual procedure becomes increasingly tedious, and the use of EDP should be considered.

Arrange the intercorrelation matrix with the criterion, 0, to the right, as shown below. The order of the predictors does not make any difference.

Intercorrelation Matrix

	1	2	3	4	0
1	1.0000	.5600	.4900	.4200	.5500
2		1.0000	.6300	.4600	.5300
3			1.0000	.3900	.5200
4				1.0000	.6400
0					1.0000

The complete solution with checks is shown in Table 1.

Table 1

Complete Solution of Multiple Correlation and Beta Weights

	1	2	3	4	0	β
1	1.0000	.5600	.4900	.4200*	.5500*	.2299
2		1.0000	.6300	.4600	.5300	.0863
		.6864	.3556	.2248*	.2220	
			.5181	.3275	.3234*	
3			1.0000	.3900	.5200	.1846
			.5757	.0677*	.1355	
				.1176	.2354*	
4				1.0000	.6400	.4318
				.7420	.3204	
					.4318	
0					1.0000	
					.4555 = $1_2 - R^2$	
					R^2 = .5445	
					R = .7379	

Solution for betas:

$\beta 4 = .4318$

$\beta 3 = .2354 - (.4318)(.1176) = .1846$

$\beta 2 = .3234 - (.3275)(.4318) - (.5181)(.1846) = .0863$

$\beta 1 = .5500 - (.4318)(.4200) - (.1846)(.4900) - (.0863)(.5600) = .2299$

$$R^2 = \Sigma B_i r_{0i} = (.4318)(.6400)+(.1846)(.5200)+(0863)(.5300) +(.2299)(.5500) = .5445$$

Final check with normal equations:

$$\Sigma B_i r_{1i} = r_{10} = (.2299)(1.0000)+(.0863)(.5600)+(.1846)(.4900) +(.4318)(.4200) = .5500$$

$$\Sigma B_i r_{2i} = r_{20} = (.2299)(.5600)+(.0863)(1.0000)+(.1846)(.6300) +(.4318)(.4600) = .5300$$

$$\Sigma B_i r_{3i} = r_{30} = (.2299)(.4900)+(.0863)(.6300)+(.1846)(1.0000) +(.4318)(.3900) = .5200$$

$$\Sigma B_i r_{4i} = r_{40} = (.2299)(.4200)+(.0863)(.4600)+(.1846)(.3900) +(.4318)(1.0000) = .6400$$

The worksheet is set up in blocks of three rows each, except the first block, which consists of only one row. The top row in each block consists of the zero order correlations from the intercorrelation matrix, with unity in the diagonals. In Table 1, these entries are underscored. The work is completed block by block, working down the worksheet. With the completion of each block, one variable is eliminated.

The second row of each block is obtained by subtracting from each zero order r in the first row a number of products equal to the number of preceding or higher blocks in the worksheet. In Block 2, one product is subtracted from each zero order r. In Block 3, two products are subtracted, etc. The maximum number of subtractions required is performed in the last block and equals the number of predictors in the problem. Thus, with nine predictors, one subtraction is made from nine zero order r's in Block 2, two subtractions are made from eight zero order r's in Block 3, etc., and nine subtractions are made from one r in the last block. It is for this reason that about ten variables are suggested as a reasonable upper limit for this method. It is the accumulation of these negative multiplications in the calculating machine which affords the principal economy of the method. Only the final residual is recorded in row 2 of each block.

The products to be subtracted to obtain row 2 of each block at first appear difficult to describe, but a pattern emerges which makes the operation easy to memorize. To illustrate this pattern, consider the residual of r_{40}, .3204. The products to be subtracted from .6400 are:

Block 1: (.4200)(.5500)
Block 2: (.2248)(.3234)
Block 3: (.0677)(.2354)

These values are asterisked in Table 1. It can be seen that the first number in each block comes from row 2 of the column above the diagonal of Block 4. The second is the row 3 entry in the column above the initial zero order r. Thus, .64000000 - (.4200)(.5500)-(.2248)(.3234)-(.0677)(.2354) = .32036310, or .3204.

The exact description of what each intermediate number in the worksheet represents is beyond the purpose of the present paper. The reader who is interested in further details of multiple regression analysis is referred to DuBois (1957, 1965).

The present problem involves no negative numbers, but if negative values appear in the solution, the operator must keep his signs straight. The solution is begun in row 2, Block 2. The first entry is

$$1.00000000 - (.5600)^2 = .6864$$

The rule applies here because only one column is involved and Block 1 has only one row. The remaining entries are

$$.63000000 - (.5600)(.4900) = .3556$$
$$.46000000 - (.5600)(.4200) = .2248$$
$$.53000000 - (.5600)(.5500) = .2220$$

The third row in each block is obtained by dividing each element in row 2 by the diagonal. Thus in Block 2,

$$\cdot\frac{3556}{6864} = .5181,$$

$$\cdot\frac{2248}{6864} = .3275, \text{ and}$$

$$\cdot\frac{2220}{6864} = .3234$$

Proceeding to Block 3, the row 2 entries are

$$1.00000000 - (.4900)(.4900)-(.3556)(.5181) = .5757$$
$$.39000000 - (.4900)(.4200)-(.3556)(.3275) = .0677$$
$$.52000000 - (.4900)(.5500)-(.3556)(.3234) = .1355.$$

Again, divide the row 2 entries by the diagonal:

$$\cdot\frac{0677}{5757} = .1176, \text{ and}$$

$$\cdot\frac{1355}{5757} = .2354$$

Proceeding to Block 4, row 2 =

$$1.00000000 - (.4200)(.4200) - (.2248)(.3275) - (.0677)(.1176) = .7420$$

$$.64000000 - (.4200)(.5500) - (.2248)(.3234) - (.0677)(.2354) = .3204,$$

and row 3 = $\frac{.3204}{.7420} = .4318$ which is B4.

Repeating the row 2 operation in Block 0,

$$1.00000000 - (.5500)(.5500) - (.2220)(.3234) - (.1355)(.2354) - (.3204)(.4318) = .4555,$$

which is $1 - R^2$. R^2 therefore = .5445, R = .7379.

The solution for betas is shown at the bottom of Table 1, and follows a pattern in row 3 of each block. From the entry in Column 0 is subtracted the products of the betas and the row 3 entry in the column of the same number as the beta. The betas are entered opposite the r_{0i} correlations for convenience in computing $R^2 = \Sigma \beta_i r_{0i}$. In the present solution this sum = .5445, which checks exactly with the R^2 obtained from the last step in the table.

The normal equations provide a final check for the solution. In solving the normal equations, it is convenient to read down each column to the diagonal, and then to the right, accumulating products in the machine. In the present example, all of the checks are exact to four places.

The present discussion has been limited to a rapid and simple solution of multiple R and the beta weights. The method is efficient and accurate for up to about ten predictors. If the researcher desires additional solutions with one or more predictors omitted, the method described will provide these solutions with no great expenditure of time or effort.

References

DuBois, P. H. Multivariate correlational analysis. Harper, 1957.

DuBois, P. H. An introduction to psychological statistics. Harper & Row, 1965.

Thorndike, R. L. Personnel selection. Wiley, 1949.

Note Concerning Correlations with Weighted Composites

John S. Brand

U. S. Army Enlisted Evaluation Center

Guilford (1965, App. A) gives the formula for the correlation of one variable, o (a criterion, for example), with the sum of differentially weighted variables as

$$r_{oc} = \frac{\Sigma\, w_i r_{oi} \sigma_i}{\sqrt{\Sigma\, w_i^2 \sigma_i^2 + 2\,\Sigma\, r_{ij} w_i w_j \sigma_i \sigma_j}} \qquad (1)$$

(where $i < j$)

and the correlation between two weighted composites as

$$r_{c1c2} = \frac{\Sigma(w_i \sigma_i \Sigma\, r_{iu} w_u \sigma_u)}{\sqrt{\Sigma\, w_i^2 \sigma_i^2 + 2\,\Sigma\, r_{ij} w_i \sigma_i w_j \sigma_j}\ \sqrt{\Sigma\, w_u^2 \sigma_u^2 + 2\,\Sigma\, r_{uv} w_u \sigma_u w_v \sigma_v}} \qquad (2)$$

(where $i < j$ and $u < v$)

The weights in these formulas are raw score weights (b coefficients), since the formulas include variances. If the weights were obtained from multiple regression analysis, they were obtained from standard partial regression coefficients (β coefficients). The relationship between b coefficients and β coefficients is expressed by

$$b_i = \frac{S_1}{S_i} \beta i$$

Where:

S_1 = The unbiased estimate of the standard deviation of the variable used for standardization

S_i = The unbiased estimate of the standard deviation of variable i.

Thus $w_i\sigma_i = b_i S_i = \beta_i \frac{S_1}{S_i} S_i = \beta_i S_1$

Substituting in both equations, the S_1 value, which is a constant, can be moved to the front of the summation signs, and cancels out in both equations.

Formulas (1) and (2) can thus be expressed in terms of β weights and intercorrelations as follows:

$$r_{oc} = \frac{\Sigma \beta_i r_{oi}}{\sqrt{\Sigma \beta_i^2 + 2 \Sigma_{(i<j)} r_{ij} \beta_i \beta_j}} \qquad (3)$$

$$r_{clc2} = \frac{\Sigma \beta_i (\Sigma r_{iu} \beta_u)}{\sqrt{\Sigma \beta_i^2 + 2 \Sigma_{(i<j)} r_{ij} \beta_i \beta_j} \sqrt{\Sigma \beta_u^2 + 2 \Sigma_{(u<v)} r_{uv} \beta_u \beta_v}} \qquad (4)$$

For computing correlations with weighted composites, formulas (3) and (4) are considerably simpler than formulas (1) and (2), and reduce the rounding error incurred by the use of b weights and standard deviations.

Reference

Guilford, J. P. Fundamental Statistics in Psychology and Education. (4th ed.) McGraw-Hill, 1965.

CROSS VALIDATION OF MOS CODE 23G40 HERCULES FIRE CONTROL MECHANIC

Paul P. Foley

This paper was prompted by an unpublished article entitled "Validity, Reliability, Baloney, and a Little Mustard," which was presented by Robert E. Chandler at the 1964 convention of the American Psychological Association.

An earlier study by Cureton, 1950, had called attention to the fact that to test more than one hypothesis on a given sample renders tables of significance practically useless since the basic premise of randomness upon which the tables are built is violated.[1]

Chandler, 1964, using standard cross validation procedures, randomly selected 100 pairs of validation and cross validation groups from a sample of 134 subjects. First, he examined the nature and amount of shrinkage which he defined as the difference between the multiple R for the validation group and the weighted validity coefficient for the cross validation group. Secondly, he attempted to determine if the mean weighted validity coefficient for the cross validation sample was more accurate than the estimate obtained through the classical correlation approach using the entire sample.

He found that the multiple R for the primary group when less than the multiple correlation coefficient (R_t), computed in the conventional manner, yielded holdout correlations greater than the R_t and vice versa.

The mean of the correlations for the holdout group was taken as the best single estimate of the validity coefficient. Only 60% of the holdout correlations were able to predict R_t as accurately as $\bar{r}_b$. He concludes that the use of cross validation in a situation not requiring it is ridiculous.

Purpose

This study was undertaken to duplicate Chandler's study using a similar testing situation. The amount of shrinkage resulting from cross validation was computed and examined. The shrunken estimate was investigated to determine if it was more accurate than the estimate obtained through the conventional correlation approach using

[1]Chandler, R. E., 1964

the entire sample. A correction formula was also applied to the multiple R for the validation group to examine the nature of this type shrinkage.

Procedure

The US Army Enlisted Evaluation Center (USAEEC) routinely uses two evaluative instruments, the commander's evaluation report (CER) and the evaluation test (ET) in its annual assessment of MOS proficiency. The MOS chosen for validation is 23G40, Hercules Fire Control Mechanic.

Three coworker ratings were gathered on each of the 104 subjects comprising the validation group. The sum of these ratings were used as an independent measure of job proficiency. A one-way analysis of variance, Winer, 1962, was used to estimate the reliability of these ratings. In addition, an F test was performed to determine the significance of the reliability coefficient.

CER and ET scores for the 104 subjects were correlated with the criterion using routine multiple correlation techniques. The resulting correlation coefficient was labeled R_t.

The total sample of 104 subjects was randomly divided into two groups; and this procedure was repeated 100 times. Weights were developed for Group I and a multiple correlation coefficient (R_a) was computed. These weights were applied to the cross validation group, (Group II), and a r_b was developed. The difference between (R_a - r_b) is the amount of shrinkage as defined by Chandler. The distribution of (R_a) was developed to show the relationship between values for R_a and the shrinkage associated with the cross validation sample.

The values of r_b were plotted to check the accuracy of the shrunken estimate r_b in predicting R_t. As in Chandler's study, R_t and r_b were compared to $\bar{r}_b$ and accuracy was defined as the minimum distance from $\bar{r}_b$.

In addition to shrinkage as previously defined, a shrinkage formula was applied to values of R_a to study the distribution of the resulting c^r.

The shrinkage of R_a, indicated by the symbol c^r, was computed using the following formula in McNemar (1962):

$$c^r = \sqrt{1 - (1 - R_a^2)\ \frac{N-1}{N-3}}$$

Where: R_a: is the multiple correlation coefficient resulting from the correlation

of the CER and ET with a criterion.

N : is the size of the validation sample.

The distribution $c^r - r_b$ was developed for each value of R_a to examine and evaluate the amount of shrinkage associated with values of R_a.

Results

The reliability coefficient of .42 was significant at the .01 level as can be seen in the ANOVA table presented below:

	Sum of Squares	Df	Mean Square	F
Between	361.64	103	3.5110	1.7278
Within	422.67	208	2.0321	
Total	784.30	311		

$$f\ .99\ (103, \infty) = 1.35$$

The results from the 100 replications are shown in Table 1. N_a and N_b indicate the sample size of the validation and hold-out groups respectively.

R_a is the multiple correlation coefficient for the validation group while r_b is the correlation for Group II developed using weights from the validation group.

c^r is the shrunken multiple correlation coefficient of R_a computed using the shrinkage formula presented earlier.

Negative shrinkage values in the $R_a - r_b$ column indicate R_a less than r_b. This occured in 43% of the cases as opposed to 22% in Chandler's study. Prior to Chandler's study it was thought that weights developed on one sample and applied to another would always result in a lower correlation. Large positive values for $R_a - r_b$ were associated with large values for R_a, while negative values were associated with low values for R_a. The shrinkage was smallest near the mean of the $f(R_a)$ distribution. The difference in R_t and $\bar{R}_a$, is .02. Figure 1 shows the distribution of R_a along with the average shrinkage associated with the R_a values in each class.

The values for $c^r - r_b$ are also plotted on Figure 1, for each value of R_a. The higher values of R_a yielded high positive differences while lower values near the $\bar{R}_a$ yielded progressively greater negative values. Small values of R_a

resulted in increasingly negative values of $c^r - r_b$.

Figure 2 shows the distribution of r_b with the $\bar{r}_b$ and R_t superimposed. The difference between R_t and $\bar{r}_b$ is shown as $\pm$ .02 on either side of $\bar{r}_b$. This accounts for 14% of the area under the distribution and indicates that 86% of the values of r_b were less accurate measures of their own mean than was R_t.

Comparison of $\bar{r}_b$ and $c^{\bar{r}}$ on Table 1 indicates that the mean shrunken multiple R provides as accurate an estimate of R_t as $\bar{r}_b$.

Conclusions and Discussion

As pointed out by Chandler, the validation and cross validation of a battery of tests using the original validation sample is ridiculous.

The correlations obtained from the primary group when less than the multiple correlation obtained using the total sample resulted in holdout correlations greater than R_t and vice versa.

The best estimate of the validity coefficient if cross validation were applicable would be the mean of the holdout correlations $\bar{r}_b$. The difference between the multiple correlation using the total sample and the mean of the holdout correlations is .02. Only 14% of the r_b values were within $\pm$ .02 points of the $\bar{r}_b$.

The difference between the shrunken multiple correlation $c^r - r_b$ revealed that c^r was an overestimate of r_b in 52 cases and an underestimate in 48 cases.

This study provides further evidence for the point made by Cureton and Chandler that to use the same sample to test more than one hypothesis is ridiculous because the basic premise of randomness is violated.

Table 1

Replication	Primary Group N_a	Primary Group R_a	Hold Out Group N_b	Hold Out Group r_b	$R_a - r_b$	c^r	$c^r - r_b$
1	41	.74	63	.22	.52	.72	.50
2	46	.71	58	.21	.50	.69	.48
3	52	.66	52	.25	.41	.64	.39
4	51	.66	53	.34	.32	.64	.30
5	51	.64	53	.37	.27	.62	.25
6	49	.64	55	.19	.45	.62	.43
7	42	.64	62	.25	.39	.61	.36
8	46	.64	58	.34	.20	.61	.27
9	52	.63	52	.26	.37	.62	.36
10	48	.63	56	.31	.32	.60	.29
11	43	.62	61	.35	.27	.60	.25
12	53	.62	51	.31	.21	.60	.29
13	42	.61	62	.29	.32	.58	.29
14	43	.61	61	.32	.29	.58	.26
15	51	.59	53	.30	.29	.57	.27
16	61	.58	43	.31	.27	.56	.25
17	45	.58	59	.29	.29	.55	.26
18	52	.57	52	.40	.17	.55	.15
19	52	.57	52	.39	.18	.54	.15
20	50	.57	54	.34	.23	.54	.20
21	51	.56	53	.27	.29	.54	.27
22	49	.56	55	.32	.24	.53	.21
23	44	.54	60	.37	.17	.51	.14
24	58	.54	46	.35	.19	.52	.17
25	44	.54	60	.41	.13	.51	.10
26	53	.54	51	.57	-.03	.51	-.06
27	61	.54	43	.30	.24	.51	.21
28	43	.53	61	.42	.11	.50	.08
29	49	.53	55	.39	.14	.50	.11
30	49	.53	55	.44	.09	.50	.06
31	46	.53	58	.33	.20	.49	.16
32	50	.53	54	.37	.16	.50	.13
33	20	.52	54	.41	.11	.49	.08
34	46	.52	58	.44	.08	.49	.05
35	45	.51	59	.44	.07	.48	.04
36	49	.51	55	.40	.11	.48	.08
37	52	.51	52	.44	.07	.48	.04
38	57	.51	47	.31	.20	.48	.17
39	55	.50	49	.39	.11	.47	.08

Replication	Primary Group N_a	Primary Group R_a	Hold Out Group N_b	Hold Out Group r_b	$R_a - r_b$	c^r	$c^r - r_b$
40	53	.50	51	.40	.10	.47	.07
41	58	.50	46	.39	.11	.47	.08
42	53	.50	51	.46	.04	.47	.01
43	53	.49	51	.29	.20	.46	.17
44	63	.48	41	.46	.02	.45	-.01
45	48	.48	56	.45	.03	.45	.00
46	45	.48	59	.47	.01	.44	-.03
47	63	.48	41	.47	.01	.45	-.02
48	53	.48	51	.52	-.04	.44	-.08
49	61	.48	43	.47	.01	.45	-.02
50	51	.48	53	.39	.09	.44	.05
51	52	.47	52	.47	.00	.44	-.03
52	63	.47	41	.43	.04	.44	.01
53	42	.47	62	.43	.04	.42	-.01
54	55	.46	49	.47	-.01	.43	-.04
55	41	.46	63	.46	.00	.42	-.04
56	56	.46	48	.57	-.11	.34	-.23
57	56	.45	48	.38	.07	.42	.04
58	64	.45	40	.44	.01	.42	-.02
59	49	.45	55	.48	-.03	.41	-.07
60	47	.45	57	.50	-.05	.40	-.10
61	62	.45	42	.33	.12	.42	.09
62	57	.45	47	.66	-.21	.41	-.25
63	52	.44	52	.49	-.05	.40	-.09
64	54	.44	50	.45	-.01	.40	-.05
65	53	.44	51	.43	.01	.40	-.03
66	53	.44	51	.28	.16	.40	.12
67	56	.43	48	.52	-.09	.40	-.12
68	48	.43	56	.45	-.02	.39	-.06
69	52	.43	52	.50	-.07	.39	-.11
70	57	.43	47	.54	-.11	.39	-.15
71	55	.43	49	.48	-.05	.41	-.07
72	44	.43	60	.48	-.05	.38	-.10
73	65	.42	39	.53	-.11	.39	-.14
74	49	.42	55	.51	-.09	.37	-.14
75	46	.42	58	.48	-.06	.37	-.11
76	61	.42	43	.45	-.03	.38	-.07
77	61	.41	43	.54	-.13	.38	-.16
78	66	.41	38	.58	-.17	.37	-.21
79	58	.40	46	.57	-.17	.36	-.21
80	49	.40	55	.41	-.01	.35	-.06
81	56	.39	48	.49	-.10	.34	-.15
82	49	.39	55	.51	-.12	.34	-.17
83	61	.38	43	.60	-.22	.34	-.26
84	54	.38	50	.56	-.18	.33	-.23

Replication	Primary Group N_a	Primary Group R_a	Hold Out Group N_b	Hold Out Group r_b	$R_a - r_b$	c^r	$c^r - r_b$
85	54	.37	50	.59	-.22	.33	-.26
86	59	.37	45	.62	-.25	.32	-.30
87	44	.35	60	.53	-.18	.29	-.24
88	49	.35	55	.51	-.16	.29	-.22
89	58	.35	46	.58	-.23	.29	-.29
90	47	.33	57	.53	-.20	.26	-.27
91	53	.32	51	.58	-.26	.26	-.32
92	57	.31	47	.59	-.28	.25	-.34
93	54	.31	50	.42	-.11	.53	.11
94	53	.29	51	.56	-.27	.22	-.34
95	59	.29	45	.55	-.26	.23	-.32
96	52	.28	52	.52	-.24	.21	-.31
97	57	.26	47	.66	-.40	.19	-.47
98	57	.26	47	.65	-.39	.19	-.46
99	55	.25	49	.52	-.27	.17	-.35
100	46	.24	58	.45	-.21	.11	-.34

$\bar{R}_a = .48$

$\bar{r}_b = .44$

$R_t = .46$

$c^{\bar{r}} = .44$

FIGURE 1. DISTRIBUTION OF R_a, $R_a - r_b$, AND $cr - r_b$

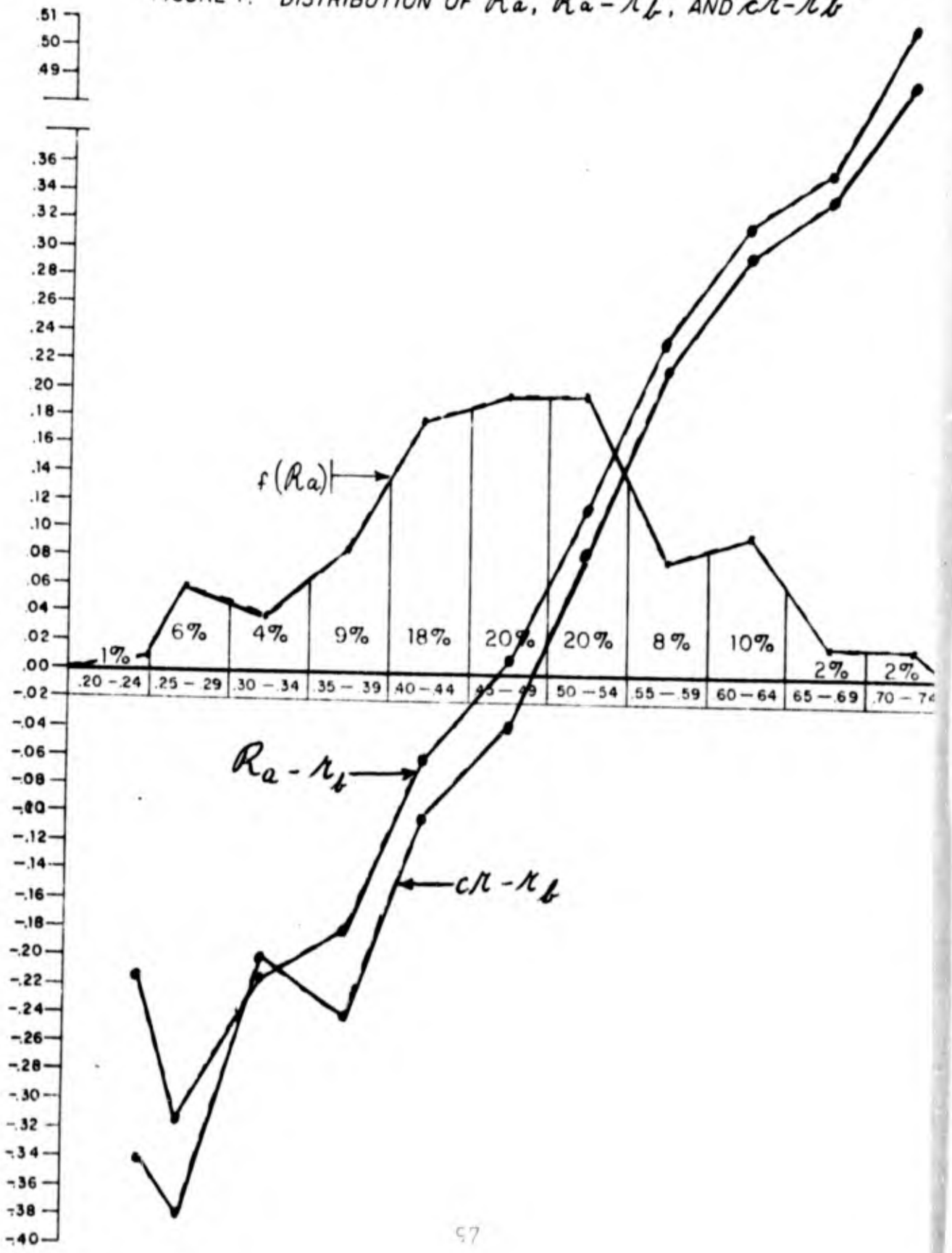
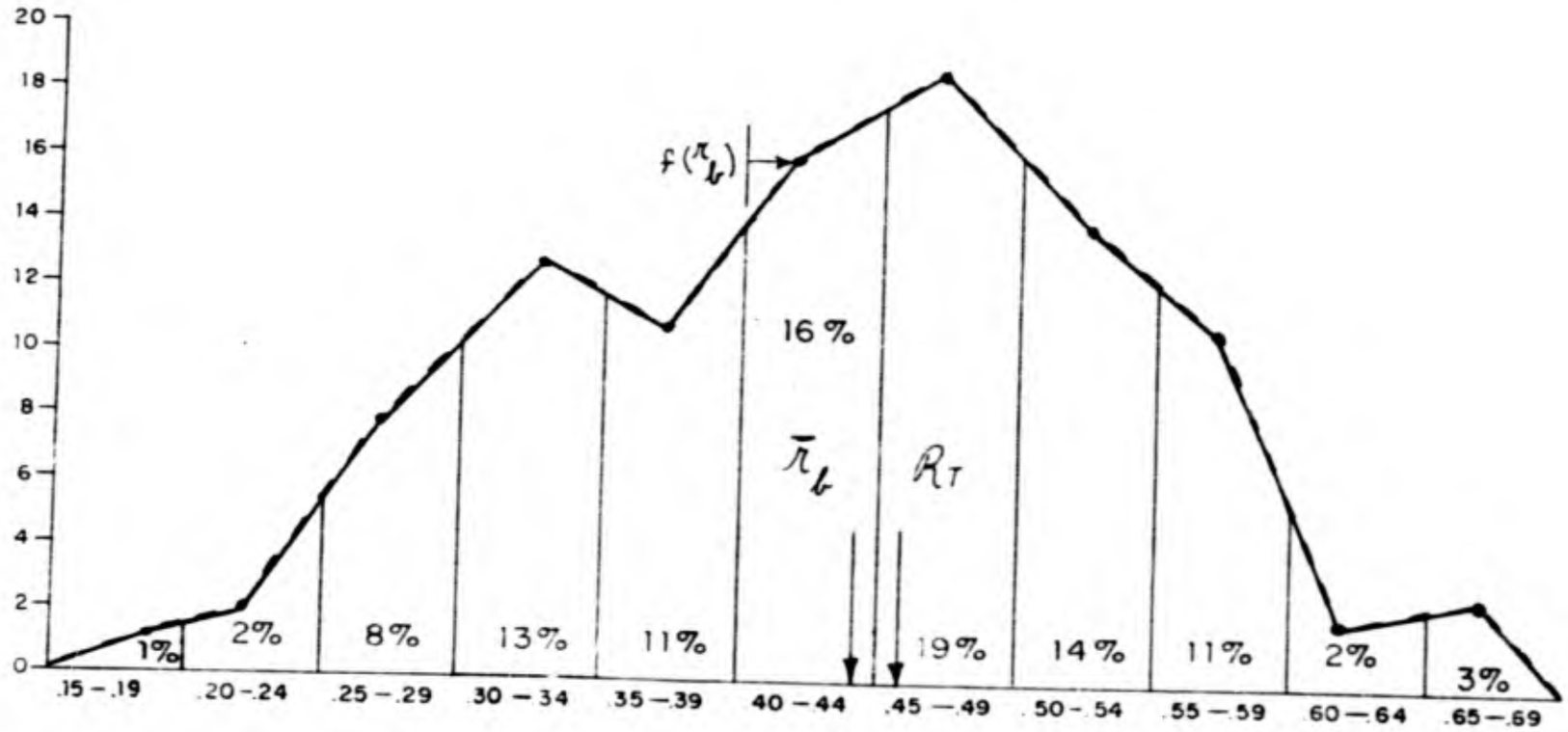


FIGURE 2. DISTRIBUTION OF π_b



References

Chandler, Robert E. Validity, reliability, bolcney - and a little mustard. General Motors Employee Research Report #8, Study #427, April 1964.

Cureton, E. E. Validity, reliability and boloney. Educ. psychol. Measmt, X (1950), 94-96.

McNemar, Quinn Psychological statistics. John Wiley and Sons, Inc., New York and London, 1962.

BLANK PAGE

MEASURING TROUBLESHOOTING SKILLS OF MAINTENANCE PERSONNEL BY PAPER AND PENCIL TESTS

John L. Finucane
U. S. Army Enlisted Evaluation Center

GENERAL

During the winter and spring of 1966, the logistics problem of the US Army in Vietnam became critical. The problem was critical because spare parts for repairing combat, communications, and avionics equipment were in short supply. The shortage of replacement parts resulted from the fact that many millions of dollars of combat, communications, and aircraft avionics parts were piled up on repair lines, waiting for repairs. Most of the available spare parts were used up as replacement items or were back-ordered to replace parts on the repair line. A serious threat to combat effectiveness existed in Vietnam.

This was a bad situation. The situation was considered worse, however, when it was determined that many of these parts waiting for repairs were (1) not defective at all, or (2) had only minor defects which should have been corrected by operators or lower category maintenance personnel. Paper work accompanying the parts did not always correctly state the defect existing in the part.

Unsatisfactory diagnosis of malfunction was the major cause of this huge backlog on the repair line. This suggested that operating and lower echelon maintenance personnel were not properly diagnosing minor equipment malfunctions. This created questions as to the training of operators and lower level maintenance personnel. Questions were also raised concerning the method of qualifying such personnel in their MOS. This last question related directly to the USAEEC, which supervises development of tests used, in part, for determining if enlisted personnel are MOS qualified.

A board of inquiry was established by the Chief of Staff of the US Army to study this problem. Lt General Frederick J. Brown was recalled to active duty to head this board and find means of solving the problems involved.

PURPOSE OF THE MOS EVALUATION TESTS

The purpose of MOS evaluation tests is to evaluate the individual's knowledge and skills in a particular MOS and skill level. (AR 600-200, para 1 - 2x). The Commander's

Evaluation Report is designed principally to evaluate the individual's performance of duty in a particular MOS. Where feasible, performance tests are administered to evaluate the individual's ability to perform a particular skill. (AR 600-200, para. 1-2 o). Scores from these three tests are statistically combined to obtain the evaluation score. It an individual's evaluation score is 70 or higher, this "certifies" (denotes) he is qualified to perform duty satisfactorily in the primary MOS and skill level in which he is evaluated (AR 600-200, para. 1-2ab).

Evaluation tests attempt to evaluate an individual's knowledge of troubleshooting, adjustment, and repair procedures and techniques required for common and unusual equipment malfunction, where this job information is required of equipment operators and maintenance personnel.

TROUBLESHOOTING QUESTIONS

The scientific method is employed in adjusting, troubleshooting, and repairing equipment. First, the problem, or malfunction, is studied and defined, and the possible solutions to the malfunction are listed. The most probable solution is tested for feasibility; if the most probable cause of the difficulty is not the solution, the next most probable cause or solution is tried, etc., until the cause of the malfunction is identified and eliminated.

The scientific method of solving or diagnosing equipment malfunction is incorporated in Army MOS evaluation tests. Evaluation test outlines for operating and maintenance MOS usually include diagnosis of equipment malfunctions under the broad subject-matter area of "troubleshooting of equipment". This area should contain questions requiring knowledge of the causes of equipment maladjustments or malfunctions. For skill levels below 40 (leader), the questions probably should be confined to the causes of common maloperation or malfunctions each skill level or pay grade is expected to know; for the leader (40) and supervisor (50) skill levels, the questions should include both common and unusual causes of equipment malfunction or maloperation.

Troubleshooting questions used in Army MOS evaluation tests take the following general patterns:

Project verbal questions on screen.

EXCESSIVE CARBON DEPOSITS IN A 4-STROKE CYCLE ENGINE COMBUSTION CHAMBER WILL DECREASE THE

A. C.
B. D.

WHAT EFFECT DOES POWER OVERLAP HAVE ON ENGINE OPERATION?

A. C.
B. D.

WHAT IS THE MOST PROBABLE CAUSE IF THE LANDING GEAR ON THE CV-2 (CARIBOU) FAILS TO LOWER?

A. AIR IN THE INTERLOCK PASS
B. AIR IN THE HYDRAULIC RAM
C. A SHORT IN THE SOLENOID COIL
D. A SHORT IN THE DEPRESSOR KEY

WHAT IS THE NEXT MOST PROBABLE CAUSE IF THE LANDING GEAR ON THE CV-2 (CARIBOU) FAILS TO LOWER?

A. AIR --- PASS
B. AIR --- RAM
C. AIR --- COIL
D. A --- KEY

The above questions are totally verbal.

Troubleshooting questions using graphic illustrations are also incorporated into our tests. Actual malfunctions of equipment, such as measuring instruments, electrical and electronic circuits, gears, etc., are illustrated in two-tone photographs and straight-line drawings. Questions derived from these illustrations require knowledge of causes of malfunctions shown. Examples of illustrated questions follow:

Project illustrated questions on screen.

1. WHAT IS THE "D" WINDING IN FIGURE_____?

A. C.
B. D.

2. WHAT IS THE FUNCTION OF THE "D" WINDING IN FIGURE _____?

A. C.
B. D.

3. WHAT RESULTS IF THE "D" WINDING IN FIGURE _____ IS SHORTED?

A. C.
B. D.

In conducting their inquiry the Brown Board developed some performance tests designed to measure proficiency in diagnosing actual malfunctions, that is, ability to troubleshoot in certain specific situations. These tests were administered to samples of soldiers holding the MOS in which troubleshooting is considered an essential part of the job. Another aspect of the Brown Board study was to examine some of our MOS evaluation tests and to identify the troubleshooting questions contained in them. These questions were

not necessarily the same troubleshooting situations as were performance tests. They do, however, test troubleshooting information. The soldiers to whom the troubleshooting performance tests were administered had also been given our MOS evaluation tests as a part of the Army's evaluation program. Sufficient information from the performance tests and the MOS evaluation tests were available for the information presented in this paper. The following data were obtained by correlating individual soldier's scores on "diagnosis" questions with their CER, raw evaluation tests (ET), and composite (ET + CER) evaluation scores.

The correlations obtained are as follows:

MOS	NUMBER OF QUESTIONS	NUMBER OF EM	CORRELATIONS	
45G20	23	7	CER vs Diagnosis Questions	r= .23
			ET vs Diagnosis Questions	r= .99**
			ES vs Diagnosis Questions	r= .96**
			CER vs ET	r= .17
45G30	22	10	CER vs Diagnosis Questions	r=-.01
			ET vs Diagnosis Questions	r= .77**
			ES vs Diagnosis Questions	r= .51
			CER vs ET	r=-.57
63C20	23	49	CER vs Diagnosis Questions	r= .19
			ET vs Diagnosis Questions	r= .58**
			ES vs Diagnosis Questions	r= .33*
			CER vs ET	r= .18

(r's are within rounding error of those reported by the "Brown Committee".)

Project data on screen.

The performance tests, in the case of MOS Code 63C20, consisted of eight cells, each cell containing an engine and appropriate technical manuals pertaining to the engine. One cell had an engine with a bad spark plug; another had an engine with a defective carburetor; another a defective starter; etc. Correlations were obtained between the individual's score on this performance test and the individual's ET, CER, ES, and diagnosis question scores. The following data were obtained from these comparisons:

Project performance test data on screen.

PERFORMANCE TEST vs ES	r = .19
PERFORMANCE TEST vs ET	r = .31*
PERFORMANCE TEST vs CER	r = .14
PERFORMANCE TEST vs DIAGNOSIS QUESTIONS	r = .33*

CONCLUSIONS

In general, no positive conclusions for all MOS evaluation tests can be drawn from the above data due to the small number of soldiers and MOS from which data were derived. (There are over a million soldiers and more than 900 MOS skill levels in the Army). Yet some definite trends are apparent:

1. Consistently higher correlation between diagnosis questions and MOS evaluation test scores for individuals than for CER or total evaluation scores. These correlations are statistically significant as indicated above.

2. Consistently low correlations between scores made by individuals on CER when compared with individual scores on Diagnosis Questions and when CER score is compared with Brown Board Performance Tests.

3. Slightly higher correlation between diagnosis questions and performance test scores than between performance tests and ET scores made by individuals.

It is assumed that the performance test used by the Brown Committee is valid, i.e., the results (scores) measure the actual ability of soldiers to diagnose equipment malfunction. Thus the slightly higher correlation between Diagnosis Questions and Performance Tests than between ET and Performance Test scores in the case of MOS Code 63C20 may prove significant. If this assumption of performance test validity is correct, then both the ET and the Diagnosis Question scores have a definite but small relationship with those soldiers most proficient in diagnosing equipment malfunctions. A higher relationship may have been obtained if the troubleshooting questions could have been developed or the same or similar situations used in the performance tests.

SUMMARY

Following the assumption that the Brown Board Performance Test was valid, data obtained from these experiments establish a trend indicating that a significant positive relationship exists between Diagnosis Questions and ET score and the soldier's Performance Test score. In addition, the ES and the CER have a slight, almost insignificant relationship with a soldier's ability to diagnose specific equipment malfunction. A slightly higher coefficient of correlation was obtained between Diagnosis Questions and Performance Test scores of individuals than between ET and Diagnosis Questions scores. This trend indicates that high individual

scores on equipment malfunction Diagnosis Questions may predict slightly more accurately an individual's ability to satisfactorily diagnose specific equipment malfunction than will the overall ET score.

If the Brown Board performance tests are valid (and they appear to be), the statistically significant correlations between scores made by individuals on performance tests and evaluation tests tend to support the view that MOS evaluation tests are valid measuring instruments of job performance.

RECOMMENDATIONS

Based upon the positive findings of this study, the following suggestions are made:

1. All "operator" (i.e., MOS requiring operators of equipment to perform user maintenance) evaluation test outlines contain one broad subject-matter area to test servicemen's knowledge of "CAUSES OF EQUIPMENT MALOPERATION" and one broad subject-matter area to test their knowledge of "USER MAINTENANCE OF EQUIPMENT" which would include cleaning, oiling, organizational maintenance adjustments, and component or part replacement, etc. where required to eliminate maloperation.

2. All maintenance and/or repair evaluation test outlines should contain broad subject-matter areas to test DIAGNOSIS OF EQUIPMENT MALFUNCTION and REPAIR/ADJUST TO ELIMINATE EQUIPMENT MALFUNCTION.

3. Correlations should be obtained between existing performance tests and diagnosis and/or repair questions and evaluation test scores for individuals, where existing.

4. Where feasible, military testing services should prepare and administer additional job-performance tests (on a one-time basis) to representative samples of soldiers working in pertinent jobs and obtain coefficients of correlation between individual evaluation test raw scores and diagnosis and/or repair question scores and the individual performance test scores.

5. Based upon the results obtained from the preceding four steps, evaluate the feasibility of using the raw ET score and/or the raw diagnosis (and repair) questions score of individuals to predict the best performers on the job.

CURRENT TRENDS IN BANDSMAN MOS PERFORMANCE TESTING

A. J. Strigari
US Army Enlisted Evaluation Center, Fort Benjamin Harrison
Indianapolis, Indiana

To adequately grasp the significance of a current trend it is oftentimes necessary to begin with a search of the past. If one wished to understand, for example, the rationale behind the first conceived psychological tests of the early nineteenth century, certain recognition would have to be given to contributing influences from the sciences of physiology, and of physics since the founders of experimental psychology were schooled in the traditions and backgrounds of these disciplines.

Today, Army bandsmen are required to take two performance tests. Specially prepared original music selections written by the US Army Element, School of Music, are printed in test booklet format and then distributed to test control officers in the field. Examinees are given a copy of one of the performance tests 30 days ahead of time. The second performance test is handed out on the day of testing. Both tests for each examinee are recorded on magnetic tape. A trained technician who is also a musician monitors the recording. The tapes are replayed to determine their adequacy and then are forwarded to a board of Army musicians who evaluate them.

The development of Bandsman MOS performance testing to its present state began in 1965 when the US Army Enlisted Evaluation Center conducted an experimental study to investigate the feasibility of incorporating a second performance test into the bandsman program. Prior to 1965, the state of bandsman performance testing was characterized by an almost exclusive concern with the sight reading ability of the Army musician. It became apparent eventually that the sight reading test alone could not provide an equitable measure of the capacity certain musically gifted individuals demonstrate for the expression of music even though they are not very good sight readers. Conversely, it was also believed that the sight reading test alone could not provide an equitable measure of those individuals who, though possessing a high degree of sight reading ability, were lacking in the capacity to express music artistically.

The prime purpose of the experimental study was to determine if a significant increase in the accuracy of prediction of an individual's consummate musical skill would result when

a second test of performance - one designed to stress the expression and interpretation of music arrived at through practice, was introduced into the performance testing program. A further purpose of the experimental study was to determine the effectiveness of a newly developed scoring form which could be used by bandsman evaluation boards in appraising the skill possessed by examinees.

The method used in the experimental study took under consideration the official sight reading performance test, a prepared performance test, co-worker ratings, the currently used scoring form and a newly developed scoring form. The sight reading performance test was recorded on magnetic tape following a 7 minute warmup and practice phase by each examinee. The advance preparation performance test was distributed to the same examinees approximately 10 days ahead of the date on which the recording would take place. The examinees participating were selected from 10 US Army bands located within the continental United States. Tapes containing the two types of performances were forwarded to the US Army Element, School of Music for scoring which was accomplished using both scoring forms. The sight reading test was scored several months ahead of the advance preparation test. At the outset of the experiment, co-worker ratings were obtained for the examinees who were to participate in the study. These ratings would serve as an external criterion against which the experimental test scores and the official sight reading scores would be compared. All data collected was then sent to the US Army Enlisted Evaluation Center for analysis. Intercorrelations using the Pearson product-moment technique were obtained among the two scoring forms, the official sight reading socres, the advance preparation experimental scores, and the co-worker ratings.

Results showed that correlation co-efficients between scoring forms and performance tests ranged from .67 to .73. Co-efficients between the criterion and the performance tests ranged from .50 to .55. The validity co-efficient between the criterion and the advance preparation test was .55. Correlations between combined test scores and the criterion revealed that a significant increase in predicting musical performance would result from the use of both performance tests.

Recommendations following upon the experimental study strongly advocated the adoption of the newly developed scoring form found to be equivalent to the present form but more economical since it required less time to complete, and the use of an advance preparation and sight reading test for the next testing session in January 1966.

In January 1966, approximately 1200 bandsmen were tested for the record. The incorporation of the advance preparation performance test into the official evaluation program represented a first in Army testing procedure in that examinees were actually allowed to retain the copy of this test until the time scheduled for recording. An interval of one month (December) was established as the period for practice between test distribution and test recording. It would be well to point out here that in the opinion of military musical educators, the condition of enabling bandsman to-practice in advance of testing closely corresponded to the rehearsal situation so commonplace in band preparation. Official test conditions had now become more analogous to the standard procedures of every day training for Army musicians. The implication of this suggests that considerable face validity had been added to the testing program for bandsman and also that the formality of the testing situation with its attendant atmosphere of apprehension for the examinee would be lessened.

The administration procedures for January 1966 were changed to include a pre-test meeting between examinees and the test administrator. During this meeting the advance preparation test was distributed to all those eligible for testing in the bandsman MOS. The practice period ensued. One month later examinees again met with the test administrator and received instruction on procedure for the official session. The sequence of testing steps included a one minute warm-up period, recording the advance preparation test, a seven minute practice period, and recording the sight reading test. Care was exercised to avoid any reference to the examinee's personal identity, skill level, length of service, or assignment location during the taping sessions. Test tapes were then sent to the appropriate bandsman evaluation board for scoring.

An analysis of the relationship of the two types of performance tests for 1966 showed a high correlation existing between them. This high relationship was not found to exist when the experimental study was conducted. A re-examination of the procedure followed in the current year indicated that the cause of the high correlation may be attributable to inherent "halo-" or rater bias on the part of the bandsman board members. It will be recalled that during the experimental study, the advance preparation test remained separated from the sight reading test because the two types of tests were administered at different times. They were also evaluated at different times. Consequently, any given examinee's performance was not present on the same tape nor were his performance tests evaluated in consecutive order.

The administration of performance tests for bandsmen in January 1967 will reflect the changes of previous years. Only through such a consistent and constant refinement process

can efforts to perfect the standards sought after in on-going human appraisal be realized.

DO SCHEMATICS HELP

By
Fred B. Honn
US Army Enlisted Evaluation Center, Fort Benjamin Harrison
Indianapolis, Indiana

INTRODUCTION

PURPOSE

The purpose of this paper is to examine the merits of using items based upon schematic illustrations and items based upon problem situations without benefit of any illustrations as currently used in the Army's MOS evaluation tests.

BACKGROUND

The basic purpose of MOS evaluation tests is to measure the job proficiency of enlisted personnel in their everyday operation on the job. Those soldiers with the highest degree of job proficiency should score highest on the MOS evaluation tests. Those with the lowest degree of proficiency should score lowest on the tests. This has been shown to be true to this extent: on approximately one-third of the tests validated against an outside criterion, the validity coefficients have been in the .40 to .60 range; also, on about three-fourths of the tests validated, the validity coefficients have been above .25.[1]

The performance of items in MOS evaluation tests is evaluated in part by the p-value (difficulty level) of the item and by the discrimination index (r_{isma}) of the item. The latter statistic (point bi-serial r) is a measure of the extent to which the item discriminates between the upper (high scoring) and the lower (low scoring) groups within a major subject-matter area of the test.

The desired range of p-values for a 125 item MOS evaluation test is between .50 and .[illegible] The minimum acceptable discrimination index is +.20. During the last several quarterly test periods it has come to the attention of the Enlisted Evaluation Center that far too many schematic items have had p-values which were below the acceptable range as well as having unsatisfactory discrimination indices. Most often these items have been too difficult and too many have had to be replaced due to an unsatisfactory index (less than .20).

[1] Thus far the MOS evaluated tests which have been validated cover about one-third of the Army's enlisted personnel who have been tested.

In the comparison of performance for two types of items in MOS evaluation testing, the better performing type would have a higher p-value (relatively easier) up to an acceptable maximum, a higher discrimination index, and fewer items would have to be replaced because of r values of less than .20.

HYPOTHESIS

This project was designed to test the hypothesis: Situational type items in Army MOS Evaluation Tests perform better (have a higher p-value and a higher r value) than schematic based items.

PROCEDURE

Since there are numerous MOS evaluation tests which do not utilize schematics, the tests used in this study were, of necessity, somewhat restricted. Therefore, the first ten MOS in three units of the Test Development Branch of the Army's Enlisted Evaluation Center were selected in numerical order. This otherwise random selection yielded sixteen evaluation tests from the ten MOS, all of which were administered during November 1965, and a total of 115 items of each type. The test population of these tests varied from a low of 26 to a maximum of 1,024. In short, the data used in this study are based upon the testing of 3,321 enlisted soldiers.

The sixteen tests were distributed among the occupational areas as follows:

OCCUPATIONAL AREA	NO. OF TESTS
Combat Surveillance + Target Acquisition	2
Guided Missile Electronic Maintenance	5
Radar + Microwave Maintenance	2
Field Communication Equipment Maintenance	1
Fixed Plant Communication Equipment Maint.	2
Armament Maintenance	1
Power Production and Distribution	1
Engineer Heavy Equipment Oper. + Maint.	2
	16

One additional point of clarification. Neither the Enlisted Evaluation Center nor this paper makes any attempt to refute Bloom's[2] categories of objectives. Rather, we encourage utilization of test items along the lines of Bloom's

[2]Bloom, Benjamin S., editor Taxonomy of Educational Objectives.

taxonomy. This paper is limited to examining the effectiveness of application, analysis, synthesis, and evaluation as exemplified in but two particular types of items namely, schematic based and situational or problem items.

Those items used in this study were paired according to broad subject-matter areas in each skill level of each MOS. The items were such as to require mental processes beyond mere recall or recognition. The final criterion of item selection was that an equal number of items, schematic and non-schematic, be selected from each outline area of each test.

RESULTS

The results of this preliminary study show quite clearly that based upon the data available, the performance of problem situational items is better than the performance of schematic based items.

TABLE 1 shows that the problem situational items functioned much better than the schematic items as to the discrimination index. Not one of the sixteen tests showed an average r-value of less than .20 for the situational items. However, nearly half of the tests showed r-values of less than .20 for schematic items. Taken item by item, the percentage of situational items per test having r-values of less than .20 exceeded twenty on just three of the sixteen tests; but, the percentage of schematic items per test, having r-values of less than .20, exceeded twenty 10 times out of 16. Normally, wouldn't you have expected just the exact reverse to be true? Remember, these soldiers have job descriptions which emphasize knowledge of and experience with schematics. Therefore, from this data one can only conclude that the discriminating power of the situational item is much greater.

TABLE 1A shows that the difficulty levels for the situational items is more satisfactory than for schematic items.

In only one instance, 45H20 - Small Missile Systems Repairman, was the mean p-value of the situational items more than five points beyond the desired range of .50 - .75. Further, this was the only test on which the situational items were significantly too difficult. On the other hand the schematic items showed a mean p-value which was more than five points beyond the range on nine of the sixteen tests: twice, the items were too easy but seven times schematic items were rather difficult.

TABLE 2 shows the performance of schematic and situational items by occupational areas. Clearly, the situational items performed better.

In six of the eight areas, the schematic items showed mean p-values below the desired range. I. just one occupational area was the mean p-value for situational items below the desired range - Armament Maintenance. Even then, the difficulty level for the situational items is more satisfactory. Further, in only two areas - Combat Surveillance and Fixed Plant Maintenance - did the schematic items have a more satisfactory difficulty level.

Regarding the mean r-values, the superiority of the situational items is even more pronounced. Three areas showed mean r-values below .20 for schematic items. None were below .20 for situational items. Only one area of the eight (Combat Surveillance) showed a more satisfactory mean r-value for the schematic items than for the situational items.

TABLE 3 shows the performance of the two types of items by MOS skill levels. Excuse the monotony, but again the schematic items had to take a back seat.

The mean p-value for the four skill levels showed only slight differences. While the two skill level showed a more satisfactory difficulty level for the situational items, the five skill level showed a more satisfactory p-value for the schematic items. Almost no difference showed for the three and four skill levels.

However, the mean r-values for the situational items were considerably better at every skill level. The values for schematic items hovered near or below the desired minimum of .20. The same data for the situational items was considerably above the minimum for each skill level.

A somewhat different picture comes from a validation study of one missile electronics MOS. The MOS did contain one major area which was devoted entirely to reading schematics. The validity coefficient for the area on schematics was a healthy .35. If we had similar studies for the ten MOS covered by this presentation, perhaps we'd have the answer to our quandary. Obviously, this presentation being limited to measures of internal consistency can't yield the data to be obtained from a validity study using an outside criterion. Hence, we can only conclude that more questions have been raised than have been answered.

Admittedly, this present study is far too limited to justify decisive action regarding use of schematic drawings. Due to present mission commitments - both at the Enlisted Evaluation Center and in the field agencies - the ideal research of this matter is not now feasible. It is hoped this presentation does point up the real need for further investigation so that we may shed more light on the problem

of using schematics in MOS evaluation testing.

Obviously, the researchers will have to visit our technical schools to analyze teaching, classroom procedures, and the emphasis placed on mastery of schematics in the classroom and on the course tests. Then, by consulting with field personnel, construct alternate forms of the test in such a way that the same question is asked verbally on one form and by reference to a schematic on the alternate form. Next, these alternate forms have to be administered to a representative sample of personnel who are working in the MOS selected for study. Having accomplished at least this much, our researchers can accumulate highly objective data which will permit definitive analysis of the problem.

Meanwhile, it is hoped you people will give serious thought to this matter. Here are a few questions which need your thinking:

1. How can we be sure that schematic items are job-oriented?
2. Do dozens of statistically questionable items justify increasing test printing costs from 50 to 75%?
3. Why don't schematic based items function better when our schools have been teaching schematics?
4. Why don't NCO's and supervisors show mastery of schematic items taken from current publications and referenced to current equipment?
5. How many schematics are needed on any one test to determine whether the examinee can read schematics?
6. How do you use schematics of 24" x 36" size when we start using the Cathode Tube?

Any answers and/or suggestions will be welcomed by the EEC.

Now, one final thought is offered. This paper did not ignore the appropriate theme of this conference. Is our evaluation effort just going on - or, is it truly on-going human appraisal? In short, then, in this day of tight budgets and ZERO DEFECTS, is our Uncle Sam getting a dollar's worth of value for our dollar's worth of evaluation effort?

BIBLIOGRAPHY

ADKINS, Dorothy C., et.al. *Construction and Analysis of Achievement Tests*, Government Printing Office, Washington, D.C. - 1947

BEAN, Kenneth L. *Construction of Educational and Personnel Tests*, McGraw Hill, New York - 1953

BLOOM, Benjamin S. (Editor) *Taxonomy of Educational Objectives*, Longmans, Green + Co., New York - 1956

CHAUNCEY, Henry and DOBBIN, John E. *Testing: Its Place in Education Today*, Harper + Row, New York - 1963

DOWNIE, N. M. *Fundamentals of Measurement* Oxford University Press, New York - 1958

MICHEELS, William J. and KARNES, M. Ray *Measuring Educational Achievement* Mc Graw Hill, New York - 1950

THORNDIKE, Robert L. and HAGEN, Elizabeth *Measurement and Evaluation in Psychology and Education*, John Wiley + Sons, New York - 1961

TABLE 1

Performance of Schematic and Situational Items - Army MOS Tests

MOS SKILL LEVELS	NR. OF ITEMS TEST	MEAN r		% ITEMS WITH r<20	
		SCHEM	SIT	SCHEM	SIT
17F20	7	.413	.243	0	43
17F40	6	.345	.245	17	17
22G20	6	.210	.385	50	17
22J20	10	.150	.310	80	10
22J40	10	.356	.359	10	0
22K20	10	.272	.339	20	10
22K40	6	.169	.335	50	17
26L20	10	.135	.294	80	0
26L40	8	.215	.369	37	0
31B20	10	.207	.398	50	0
32Z40	5	.070	.288	60	0
32Z50	5	.232	.286	40	20
45H20	4	-.027	.305	75	25
52D20	5	.246	.322	20	20
62C20	7	.313	.263	14	14
62C30	6	.147	.395	67	33

r - VALUES

TABLE 1A

P - VALUES (.50-.75)

MOS SKILL LEVELS	NR. OF ITEMS TEST	SCHEMATIC		SITUATION	
		MEAN P	BEYOND RANGE	MEAN P	BEYOND RANGE
17F20	7	77.7	+3	48.6	-1
17F40	6	84.5	+9	65.3	
22G20	6	59.7		53.3	
22J20	10	37.0	-13	60.0	
22J40	10	56.6		66.5	
22K20	10	35.7	-14	48.9	-1
22K40	6	37.2	-13	45.5	-5
26L20	10	33.9	-16	54.1	
26L40	8	50.1		61.5	
31B20	10	34.8	-15	54.1	
32Z40	5	53.7		50.8	
32Z50	5	56.0		49.0	-1
45H20	4	25.0	-25	35.0	-15
52D20	5	32.2	-18	59.8	
62C20	7	46.0	-4	54.0	
62C30	6	45.2	-5	45.0	-5

TABLE 2

Performance by Occupational Areas

OCCUPATIONAL AREAS	NR. OF ITEMS	MEAN P			MEAN r		
		SCH	SIT	DIFF	SCH	SIT	DIFF
Combat Surv. + Target Acq.	13	.797	.563	-23	.392	.244	-15
Guided Missile Electronic Maint.	42	.447	.559	11	.240	.343	+10
Radar + Microwave Maint.	18	.411	.574	+16	.170	.327	+16
Field Commun. Equip. Maint.	10	.348	.541	+19	.207	.398	+19
Fixed Plant Comm. Equip. Maint.	10	.551	.500	-5	.152	.287	+13
Armament Maintenance	4	.250	.350	+10	.027	.305	+33
Power Production + Distribution	5	.320	.598	+28	.246	.322	+8
Engr. Heavy Equip. Oper. + Maint	13	.466	.501	+4	.236	.324	+9

TABLE 3

Performance by Skill Levels

SKILL LEVELS	NR. OF ITEMS	MEAN P			MEAN r		
		SCH	SIT	DIFF	SCH	SIT	DIFF
Spec. 2	69	.420	.529	+11	.221	.320	+10
Spec. 3	6	.452	.450	-.2	.147	.395	+25
NCO 4	35	.564	.593	+3	.249	.328	+8
Sr. NCO 5	5	.564	.492	-7	.232	.286	+5

INTRA-INDIVIDUAL VARIANCE AS AN APPRAISAL TECHNIQUE (ABSTRACT)*

Claude F. Bridges
Office of Research
U. S. Military Academy

The variance of an individual's scores on a battery of measures, or major components thereof, is a commonly unused aspect of individual differences. Such "profile variance" is a readily obtained statistic reflecting one aspect of differences in the interrelationships of measures within individuals. The population variance in intra-individual variance on a group of measures often is larger in magnitude than is the population variance in the composite score based on these component measures. Appropriate formulas are given. Data presented shows that intra-individual variance, on beta weighted measures in a battery selected to predict academic success at the U. S. Military Academy has significant curvilinear relationship with both academic average and 'over-achievement." Cadets earning higher academic averages than predicted by the battery were significantly more heterogeneous, on the measures in the battery, than "parachievers" and "under-achievers." Academically successful cadets were not significantly more heterogeneous than cadets separated for academic failure.

*The complete paper with considerable additional data will be available to qualified requestors from the Office of Research, USMA, West Point, N. Y. 10996

TRIBULATIONS OF TESTING

Philip Ash, Research Assistant to the Vice President
Inland Steel Company

You may have observed, as I have, that in this changing world in which we live the spotlights move, the controversies change, the hot seats are never the same. For instance, for many years public education went its quiet way. Nobody paid much attention to what teachers and principals and superintendents and boards of education were up to -- nobody, perhaps, except the ladies of the PTA and the taxpayers associations. And then, quite suddenly, public education became the storm center over which the bitterest fights of local government and local politics raged. City superintendents had to deal with striking teachers, boycotting parents, and picketing civil rights groups. It has been too much for some of them.

The same thing has happened to real estate brokers. This profession has become a storm center, the focus for a public clash of issues and values. The policies of the Chicago Real Estate Board, for example, have become a matter of public interest and public concern. And the neighborhood real estate broker has been forced to concern himself with problems which for his entire career he may never before have had occasion to deal with.

You can think of other examples as well as I. In the last couple of years the police of our cities have become the focus for public attention and controversy. The more strongly they are criticized by some groups, the more energetically they are defended by others. No matter what they do, somebody feels that it's wrong.

Today, testing and testers are also on the griddle -- not only out here in industry, but all over. We have been receiving some of the same kind of critical attention that the press and the public have extended to educators, real estate men, and police officials. We too have become the center of a storm of controversy. Our traditional tools, techniques and theories, which for years were used with confidence and without question by schools and colleges, by employers and guidance couselors, by civilian and military agencies of government, are being subjected to critical examination by everyone from unhappy parents to the United States Senators -- especially the latter.

These criticisms arise from a variety of sources, and they have a variety of causes. Some of the critics don't know what they're talking about, but others are well-informed

and have made effective statements about the use and misuse of tests. Tonight I should like to discuss three areas of criticism of testing -- the invasion-of-privacy problem, the racial discrimination problem, and the attack on testing for employment and promotion which is being made by the unions. Because I believe most of us here tonight are more directly concerned with the use of testing in connection with matching a man with a job, I shall not go into the attack on testing in the schools. That is a subject for another night.

The first shot that I can remember in the campaign against testing was fired over a decade ago, in William H. Whyte's book The Organization Man. Whyte's blast began what might be called the anti-conformist battle against testing, later to develop into the anti-invasion-of-privacy campaign. Under-ground much of the time, occasionally it broke into public view when a journalist or other freelance writer would publish an article or a book lambasting tests and testing and condemning anything more sophisticated than a typing speed test.

I suppose all of us remember Whyte's Six Commandments for Taking Personality Tests. He advised: "To settle on the most beneficial answer to any question, repeat to yourself:

(a) I loved my father and my mother, but my father a little bit more.
(b) I like things pretty well the way they are.
(c) I never worry much about anything.
(d) I don't care for books or music much.
(e) I love my wife and children.
(f) But I don't let them get in the way of company work."

Six years later Martin L. Gross wrote a book called the Brainwatchers. This one was an attack aimed specifically at personality testing. But somehow the author's aim was not good. A fair number of reviewers seemed to feel that Gross had come out with a broadside against all types of tests and testing.

At about the same time there were at least two other books and a spate of magazine articles, all of which took as their theme the unfairness of permitting test results to decide the fates of people -- in industry or in education. At that time, no one paid much attention to you in the military or government. Your turn came later. Guerilla attacks were made by parents of public school children in a scattering of communities, and they won a big one when the public schools in New York City discontinued all types of testing -- except course testing in the classroom by the teacher.

Sometimes it was not too easy to discern exactly what the critics were upset about. Part of the time it appeared to be that the tests they were objecting to were invalid; and part of the time they seemed to be afraid that they were all too valid. It wasn't clear whether the critics were objecting to personality instruments because they sometimes didn't report the truth or because they did. At any rate, as public issues have a way of doing, the question of use of tests -- specifically personality tests -- found its way to the halls of Congress. Committee hearings were held in 1965 and 1966 by committees of both the House and the Senate; the critics and the experts were heard from, and pressure was brought on federal agencies to modify their policies or to discontinue tests use altogether. Heads of such agencies as the Civil Service Commission and the Peace Corps were asked to explain what tests were used and why, and to show that they in fact predicted job performance. Witnesses told horror tales about being questioned about their sex lives if they wanted jobs as office clerks. Sympathetic congressmen, who may have remembered what they had learned many years ago in law school about the right of privacy and the privilege against self-incrimination, tried to use these concepts to deal with issues like the proper use of the MMPI by federal agencies. It has been an interesting effort to watch. And a good number of psychologists have received their initiation into the perils of examination by the counsel of a congressional committee. It is an experience which I am sure they will not forget.

The public in general tends to lump personality tests, lie detectors and their operators, and electronic eavesdropping all together as things they don't like, for reasons they can't always articulate. Naturally we psychologists don't like to be put in the same class as the private detective with his bugged martini or his polygraph. And just as people feel justified in debugging their homes or offices or cheating the polygraph, they are quite willing to try to beat the tests.

The furor in the press about the privacy issue, however, did not amount to much until Senator Ervin and Congressman Gallagher took up the cause with committee hearings during the 1965 and 1966 sessions of Congress. Frequently legislation, which is the avowed purpose of having committee hearings, is the least of the results. Legislation was not necessary to secure concessions on the use of personality tests from the Civil Service Commission and other agencies. And legislation was not necessary to bring pressure on employers which do business with the government.

As a matter of fact, a bill (S.3703) has been submitted by Senator Ervin "to protect the employees of the Executive Branch of the United States Government in the enjoyment of their

constitutional rights and to prevent unwarranted governmental invasions of their privacy." It is interesting to note that the prohibition against requiring governmental employees or applicants to take psychological tests which inquire about personal matters is coupled with the provision restricting the use of the polygraph. The provision, as proposed, reads as follows:

> It shall be unlawful . . . to require or request, or to attempt to require or request any employee of the United States . . . or any person seeking employment in the executive branch of the United States Government, to submit to any interrogation or examination or to take any psychological or polygraph test which is designed to elicit from him information concerning his personal relationship with any person connected with him by blood or marriage, or concerning his religious beliefs or practices, or concerning his attitude or conduct with respect to sexual matters.

Apparently, in the mind of Senator Ervin -- and I believe in the minds of quite a few other persons -- there is not much difference in principle between the MMPI and the polygraph.

We as psychologists have an important responsibility. It is up to us to make sure that the public and its representatives know what the instruments we use can do and what they cannot do and that whatever legislation is enacted grows from an informed judgment about the relative values of privacy and of psychological assessment of employees and applicants.

However, I am afraid that some of us continue to pursue our efforts to assess people without realizing that some of the things we are doing are being seriously and sincerely questioned, and that others may already be of doubtful legality. Let me give you one example.

We have long known that many personal items which can be included on an employment application blank are valid predictors of success on the job, of turnover, and so forth. This knowledge has resulted in the development of the concept of biographical data as useful material for prediction of job success. Division 14 of the APA has assembled and published a long list of questions -- A Catalog of Life History Items -- which can be asked of prospective employees with the thought that members might find some of the suggestions useful and also contribute studies on the validity of some of the suggested questions.

I looked through the Catalog as I was preparing these remarks, with the thought in mind of seeing whether or not questions were included in the Catalog which might be improper under the provisions of state and federal civil rights legislation, as well as whether or not some of them might be considered to be an improper infringement on the applicant's private life.

The first one I found was Question 1 -- How old are you now? Although this question is acceptable under federal law, there are several states which prohibit employers from discriminating on the basis of age, and in some of them -- New York, for example -- this prohibition has been interpreted as meaning that it is improper for the employer to ask an applicant's age.

Questions 2 and 3 present an interesting problem. First I should point out that nowhere in the catalog is there a question asking the sex of the respondent. Therefore we presumably do not know his/her sex. Now what is the effect of asking Question 2 -- What is your height? Or Question 3 -- What is your weight? If tall people and people who weight over 175 pounds get a plus, and people under five feet seven or weighing less than 150 pounds get a minus, could these questions be construed as discriminating against women? They would appear to tend in that direction. Certainly one would need two sets of norms -- norms for male and for female applicants -- to avoid this result.

Questions are included about the applicant's religion, including these -- How often do you read the Bible? Under usual conditions how often do you attend church? Do you attend a church other than the one of which you are a member? How much of your time is devoted to religious activity? In the matter of religion, how would you classify yourself?

There are also questions about national origin, including these -- Where did you father and his ancestors come from? Where did your mother or her ancestors come from?

The use of the answers to any of these questions as factors in employment selection would clearly be a forbidden act of discrimination under the 1964 federal Civil Rights Act and under the fair employment acts of many of the states.

In addition to these questions, there are a large number of questions which employees or applicants might feel are unwarranted invasions of their personal privacy. Some of them would be prohibited by the provisions of the bill introduced this summer by Senator Ervin. Senator Ervin's bill would

prohibit, among other invasions of privacy, interrogating federal employees or applicants for federal jobs with intent to solicit information about personal relationships with relatives, or religious beliefs or attitudes, or attitudes and conduct with respect to sexual matters. Now if we look at the questions included in the Catalog of Life History Items, we find quite a few which would clearly fall under this ban. Consider these, for instance:

> Where did you get your early information about sex matters?
> How did your father feel in regard to you going to college?
> Do you feel your father has been successful?
> Who picked on you most during your youth?
> How would you characterize your home (happy, unhappy, etc.)?
> Which one of these problems has cropped up more than others with your wife?

Now, I do not mean to suggest that none of these questions is useful in predicting success or failure on a job or in a company. Some of them may be quite useful. But the fact that they may be useful becomes irrelevant if asking them should be forbidden by law.

I realize that nobody has yet told private employers that they cannot ask prospective employees questions such as these. However, in this as in other fields, the government is or can be a very influential employer. And if the Congress concludes that the federal government should not pry into the family relationships and the sexual beliefs and attitudes of its employees, how long will it take for Congress to decide that it is equally improper for private employers to do so?

Therefore, in concluding what I have to say about the privacy issue, I should simply like to suggest that the battle between the defenders of privacy and the askers of questions has passed its preliminary stages. Privacy is probably on its way to becoming a constitutional right; not to be lightly infringed.

I should now like to turn from the privacy battlefield to the civil rights-discrimination arena. Here too the askers of questions are under attack, but the nature of the attack is different. The subject matter, although sometimes it may be personality testing, is more often aptitude or intelligence or achievement testing -- something about which the defenders of privacy do not too often concern themselves. And the attack, again in contrast to the privacy campaigners are afraid of what testing can discover; the civil rights campaigners are more worried about what it cannot discover.

Again, at the beginning there were scattered skirmishes. Individual state fair employment commissions had received a few complaints that some employers were faking scores on tests given to minority applicants or were shortening the time limits, or were otherwise discriminating in their test administration practices. These practices were clearly outlawed by the respective state FEP laws, and the situation was simple. It was a matter of catching them at it, and then proving it. But this type of complaint was made relatively infrequently and the issues involved were cut and dried.

However, the situation changed radically with the celebrated Illinois case of Myart v. Motorola. I do not propose to go into great detail about that case with you. Most of us will never forget it. But I would like to point out that the question which the case brought so forcefully to the attention of Congress, employers, and the public was not the basic question in the case and was not resolved by the decision. As things turned out, the question of whether or not the test given to Leon Myart discriminated against Negroes--or against anyone else--went essentially unanswered from the beginning to the end.

You will recall that in July, 1963, Leon Myart, a Chicago Negro, applied for a job as a television phaser and analyzer at the Motorola plant in Franklin Park, a suburb of Chicago, and was given a brief intelligence test. He was not hired after being given that test, and he was not contacted later. When several days had passed, he noticed that the company was still advertising for people to fill the job for which he had applied. Feeling that his prior education and experience should have qualified him for the job, Myart filed a complaint with the Illinois Fair Employment Practices Commission, alleging that he had "passed the test" and that he had been discriminated against because of his race.

After investigation by the Commission investigator and formal hearings before a Commission hearing examiner, the examiner issued his findings and recommendations. They were to the effect that what evidence there was did not refute Myart's allegation that he had passed the test. But in addition the examiner examined the test which, he had decided, Myart had passed and found that the test discriminated against culturally deprived persons.

The hearing examiner, therefore, also directed Motorola to discontinue using the test.

It was the second part of the examiner's opinion that caused all the trouble. The Commission on review held extensive further hearings to determine whether or not it should adopt the hearing examiner's findings, and these proceedings were almost entirely concerned with whether or not employment tests in general and this test in particular discriminate against members of minority groups--that is, whether they understate the capabilities of such persons to perform the job for which they are being tested.

After the Commission's review proceedings were concluded, it sat on the Motorola case for several months, during which public and professional discussion of the test issue continued to focus on whether or not tests may themselves be discriminatory. After the general election of November 1964 the Commission issued its decision on review, in which it found that on the basis of the evidence available, Myart had passed the test. That is, the Commission agreed with Bryant on this narrow point. But it made no finding that the test itself was or was not discriminatory. And it ordered Motorola to pay Myart $1,000, but it did not require that he be hired.

Motorola appealed the case to the courts. The first trial court (the Circuit Court of Cook County) invalidated the Commission's award of $1,000, but allowed the finding of discrimination to stand. The Illinois Supreme Court looked again at the same evidence which had been analyzed by the hearing examiner, the the full commission, and the circuit court, and found that there was not sufficient evidence to permit the commission to find that Myart had passed the test. The Illinois Supreme Court's decision was handed down March 24, 1966, almost three years after Leon Myart applied for a job in Franklin Park (July 15, 1963).

I have outlined the history of the Myart case not because I want to raise old questions or open old wounds, but because I think it is ironic that the case which is responsible for a far-reaching

rethinking of the whole field of personnel testing, for the Tower Amendment to the 1964 Civil Rights Act, for pages and pages of learned writing in both psychological and law journals, and for hours and hours of discussion in meetings of psychologists and personnel men, was not decided in the end on the issue of discrimination by tests at all, but rather on the issue of whether or not there was enough evidence to justify the Commission's finding that Myart had passed the test in the face of the Company's insistence that he had failed--on the technical evidentiary question of the legal effect of the absence of the original test paper.

But, as has happened more than once in history, significant results flowed from a small detail.

Concerned employers and personnel men have conscientiously reviewed and revised their testing programs to see whether in fact the tests they use predict job performance on the jobs being tested for, and whether or not they predict equally well for white and nonwhite applicants. Some results are in from some studies of these problems which indicate that possibly the typical paper and pencil tests used by employers do not predict job performance as well for Negroes as for Whites.

Employers are beginning to realize that differences in test scores are of value only to the extent that they predict differences in job performance. The real question is: Will an applicant with a low test score perform his job less satisfactorily than one with a significantly higher score? Or is it possible that the low scorer candidate can do as well as the candidate with the higher score?

There are beginning to be a few studies reported which bear upon this point. Although the evidence tends to some extent to suggest that tests may in fact have differential validity, however, the data are still too few to permit a really informed judgment. It is probable that some tests "discriminate" (in the social sense) some of the time, but not all tests all the time.

Studies now under way may well have profound implications for all test users. Even though it may be true that a test serves as an accurate predictor of job performance for an applicant who was born into and educated in the predominant White, highly verbal culture of the United States, this does not mean that the test will be an accurate indicator of possible job success for an applicant who has not shared--or shared only partially--in this culture.

It is clear that if these differences are established, the employer who knowingly persists in using tests which have this effect is in fact discriminating against members of minority groups and I am sure that as the sophistication of FEP commission members and courts grows, they will look for this type of discrimination in the testing practices of employers with all or mostly White workforces.

The third and last campaign against tests and testing which I should like to call to your attention is the campaign being waged by organized labor. Just last week the AFL-CIO released a statement opposing the use of both aptitude and personality tests by employers. Bert Gottlieb, the federation's Director of Research, pointed out that the promoters of some tests claim to be able to determine whether a worker is liberal or conservative, pro-union, anti-union, or neutral. If these claims are true, he said, a management--without doing anything to the union that is in existence--can in the long run seriously weaken it by using the test to weed out those people they feel would be good union members.

The unions have of course traditionally been in favor of promotion within a plant on a strict seniority basis. In general, they feel that a worker earns a promotion by being there longer than anybody else. Sometimes they are willing to recognize the desirability of having a trial or probationary period after a promotion, to give the employer an opportunity to see whether or not the worker can do the job or learn it in a reasonable time. But in most cases unions do not seem to like tests.

But it is only within the last few years that they have made much of an issue of testing. Perhaps this is because with the advent of more complicated and automated equipment, more employers have sought to use tests to help them determine more accurately who among their workforces could best learn to use this equipment.

At any rate, several unions have begun seriously to question the use of tests by employers. Most arbitrators have been willing to uphold an employer's use of tests for promotion

if, in the mind of the arbitrator, the test seems to bear a reasonable relationship to the job being tested for. Employers as a rule have relied on face validity to convince the arbitrator, and most union officials are not technically equipped to argue about test validities and other such things. However, in the last several months several union spokesmen have pointed out the importance of teaching local union leaders at least enough about personnel testing so that they can ask intelligent questions. If an employer is not equipped to show that the tests he uses do in fact predict success on the jobs for which he uses them, he may begin to expect trouble from his union.

In their standard book, How Arbitration Works, Frank and Edna Elkouri list the four factors which arbitrators use to judge a specific test: (1) that it is specifically related to the job; (2) that it is "fair and reasonable," (3) That it is administered in good faith and without discrimination; and (4) that it is properly evaluated. It goes without saying that a test which cannot be defended on these bases is not doing a good job for the employer. It can be argued that when a union forces an employer to look critically at his testing program, it is doing him a favor.

Now that I have described the nature of the three attacks on testing, you may well ask, "What defenses do we have?" Should we return to the dark ages of interview and instinct? Must we give up testing?

The answer of course is no. But we must choose our ground. We must defend what is really worth defending and not waste our ammunition on the defense of testing programs which may not be worth having. And we must remember that selection, evaluation, and assessment of persons is an everyday activity in every business or other organization. The criticisms that have been leveled also at every other form of selection and evaluation, including the personal interview. It is true that the use of a test can be illegal, it can be immoral, and it can be intrusive. This is equally true of the interview. And it can be equally true of the background investigation, the reference and credit check, and the concealed observation of the employee. The problems--legal and ethical--are not with the instrument; they are with the motives and judgment and skills of the user of the instrument.

BLANK PAGE

HUMAN ENGINEERING METHODS OF APPRAISAL

Robert L. Malone, Ph.D.
Fry Consultants, Inc.
Donald O. Shepherd
U.S. Naval Examining Center

Having worked in the field of Human Engineering in the design of military systems for a number of years prior to my present position as a consulting psychologist, I have become convinced that the methods and techniques used in this sphere of applied psychology have a great deal to contribute to programs and efforts elsewhere which seek to improve human performance. In fact, I have been advocating that we rename this approach "Performance Engineering" because in its broader concept as applied in the design and development of man-machine systems, the emphasis is appropriately on all aspects of the system and its environment in which there is some concern for man and his performance. I believe renaming of this field of work would also provide the additional timely advantage of avoiding some of the humorous yet serious misunderstandings as to what Human Engineering really is. Certainly it is not the designing or remaking of people in a literal or flippant sense. And today it is certainly more than the design of dials, controls, procedure manuals, and other equipment design aspects which a few years ago characerized Human Engineering activities. In order to represent this field more accurately in terms of the total implications of man and his performance within an organized system, the Air Force has used the term,"Personnel Sybsystem Engineering." This latter term has, however, also caused some confusion on the part of those who interpret it as dealing with traditional aspects of personnel, which involves recruiting, selection, training, and supervision of people for an organization that has already been defined and for equipment which has already been designed. While I prefer the term, "Performance Engineering", I am willing to use the term, "Human Engineering" in the remainder of my presentation today--with the understanding that it includes all aspects of system design which deal with man and his performance, and further that the various aspects of human performance are considered concomitant with and in close interaction with all other elements of the evolving system. Before outlining further the methods and operations of Human Engineering, I would like to state briefly some of the major improvements that lie ahead which I believe can and should grow out of a better cross-utilization of the methods of this field and those associated with traditional psychological testing and appraisal.

These improvements include:

...Improved definition and standardization of job tasks based on a meaningful task taxonomy
...Redesign of jobs and tasks to effect better utilization of skills and abilities within all segments of the population
...More effective performance motivation

The respective roles of Human Engineering and Psychological testing methods in bringing about important developments such as those enumerated can be brought into clearer focus by reviewing the area of Human Engineering and its methods for achieving effective human performance within the systems development context. There is no reason why this approach cannot be adapted to bring about improvements in industry and government as well as those which have been achieved in information, communications, control, weapons, and space systems.

Review of Human Engineering as a systems-oriented discipline

The role of Human Engineering in system design and development is to help achieve effectiveness of the entire system so that the system can perform its required functions within specified limits of speed, accuracy, reliability, and maintainability. Human Engineering seeks also to support design objectives of minimizing skill and training requirements, costs, and developmental time. We might say, therefore, that system design effectiveness is directly proportional to speed and accuracy with which the system performs, its ability to remain in sufficient operational condition to sustain that performance, and its ability to be restored quickly to that operational state when malfunctions or degradations occur. Its design effectiveness is inversely proportional to the levels of training and skills required by it as well as developmental and operational costs and development and activation time.

Human Engineering seeks to help achieve effective systems design through a step-by-step methodology aimed at fulfilling a number of specific design objectives which must be progressively met as the system evolves.

(Slide No. 1)
These include:[1]

(1) Should there be a human component in the system?
(2) If so, what functions should the human perform?
(3) Within these functions, how should the required tasks be organized?

---

[1]Charles E. Gustafson and Marty R. Rockway, The Air Force Personnel Subsystem Concept, Aeronautical Systems Division AF Systems Com., Wright-Patterson AF Base, Ohio, 1962.

(4) How should the equipment be designed and arranged to facilitate human and system performance?
(5) How can we best protect and sustain man during the operational mission?
(6) What skills and knowledges must each individual possess to perform his tasks and functions adequately?
(7) How many men are needed to get the total job done?
(8) What training is required and where and how can it best be accomplished?
(9) What training equipment is needed and to what extent will it help develop the required operational performance?
(10) What manuals, checklists, or other publications are needed to facilitate learning and performance?
(11) How well does the completed system perform?

The methods used to resolve these questions are applied in a somewhat definite order and in parallel with the design and development of machines or equipment.

(Slide No. 2 - Briefly enumerate items next to numbered boxes)

As this slide reveals, Human Engineering activities comprise the following steps which become more and more complete as the design progresses from the development phase to production and installation phases.

(1) During systems analysis, functions must be assigned to men and to equipment in a manner that will satisfy system functional requirements and recognize the respective capabilities and limitations of men and equipment.
(2) Human task requirements must be analyzed and specified in sufficient detail to determine requirements which must be satisfied by equipment design so that total man-machine interaction and performance will meet system performance criteria.

Criteria for the design of equipment derive in part from both general and specific human considerations. General criteria deal with those design features which research and experience have shown to be necessary to promote effective performance for any system namely, design of indicators, controls, illuminations, etc. consistent with human accuracy and limitations in sensing, processing, reacting, and decision making. Reference documents and handbooks containing these

criteria are readily available. Specific criteria for equipment design evolve from the analysis of functions and tasks to be performed by the human for a particular system, and these criteria depend on the type of functions, how they were allocated, and the concept for operating and maintaining the system. For example, in processing messages in a communication system, the human may be required to compose the message by means of a typewriter input, by means of depressing alphanumeric characters which select preformatted segments of the message, or by simply keying in the type of message to be composed automatically from pre-stored data. Whichever approach is used will govern the nature and design of equipment provided for him. Further, the procedure to be followed in preparing, checking, safeguarding, and transmitting messages for the given system will govern the surface panel design of console equipment so that correct sequencing can be facilitated.

(3) The third Human Engineering activity involves specifying the criteria for equipment design and, in turn, determining what skill, knowledge, and procedural elements are imposed on the man by the type of equipment, its complexity and limitations. In order to minimize skill and training requirements as well as selection problems, it is a customary design goal to set equipment design criteria minimally at the 5th percentile. However, cost and time constraints may require compromises or tradeoffs so that skill and training requirements incorporated into the design may at times be somewhat greater than hoped for. The importance of considering carefully how man and machine impose requirements on each other can be illustrated by the following slide which shows that the equipment input and output is dependent on the sensing, processing, and reacting activities of the human operator. Similarly, the human input and output are influenced by equipment performance.

(Slide No. 3)

(4) Important next steps involve design of the various jobs to be assigned to operations and maintenance personnel and the specification of the work environment in terms of the arrangement, functionally and physically, of men and equipment, the size of quarters, illumination levels, ventilation, safety, and other considerations influencing comfort, protection, and performance effectiveness. Task analysis data derived and organized by means of work sheets is used to guide the design of jobs. The kind of information dealt with is illustrated by this sample worksheet.

(Slide No. 4 - Enumerate briefly the major jeadings of this form)

The arrangement of men and equipment in the work environment should seek to minimize the overall effort by which behavioral links or interactions between men and equipment are implemented. To achieve this, both frequency and importance of links are examined so that a diagrammatic solution can be derived similar to that achieved with sociometric analysis. The resulting arrangement is preferred which minimizes distance and effort required to achieve the required visual and motor tasks in appropriate procedural sequence. The functional relationships based on link analysis and the resulting physical layout which attempts to fulfill the results of the link analysis are illustrated in the next slide.

(Slide No. 5)

(5) To facilitate manpower planning, selection, and training, the Human Engineering program includes the preparation of a Personnel Requirements Information document which reviews task and performance requirements and standards, outlines the duties and tasks associated with each job position, identifies the number, and type of personnel deemed applicable for each position, their prerequisites and estimated training needs. Training equipment, training and performance aids, manuals and checklists are developed or specified from the same source data used to derive the personnel requirements information. The source data includes task analysis work sheets, functional flow and block diagrams, equipment design and evaluation data, system detailed performance criteria, work environment drawings, and appropriate contractual specifications. While much of the effort is analytical, mockups and other means for providing simulation and study of design alternatives are used to evolve and verify the selected results.

(6) During all phases of system design, evaluation and verification must be conducted as an ongoing activity. As the design progresses, decisions are made which may impact the entire design of the system and perhpas undermine earlier design decisions unless the verification process is carefully maintained. This evaluation deals with both the products and the processes involved in the design of the system. Has the appropriate task information been developed? Are the tasks and procedures designed to minize skill and training requirements? Is the equipment designed in a manner to facilitate rapid and accurate task performance? Have the necessary performance aids

and support equipment for operating and maintaining prime equipment been identified and developed? Is the integrated and installed system capable of meeting its specified requirements of speed, accuracy, reliability, and maintainability?

Evaluation of these questions involves a variety of methods ranging from controlled experimental study of design variables, simulation of system operation using computers, mockups, and pencil and paper methods, checklist review of design criteria, statistical computation of reliability and maintainability, and proficiency testing of individual and group performance in operating and maintaining the system. These evaluations are ultimately focused on the objective of meeting contractually specified system criteria, and the design is acceptable if these criteria are fulfilled. When the criteria are unfortuantely not fulfilled, diagnostic analysis is required to identify those aspects of the design which have either fallen short or which could feasibly be improved to get the system performance within specified limits.

Challenges that lie ahead

As technology increases and more complex systems and sophisticated equipments are developed, we can anticipate further challenges in the effective development and utilization of available manpower to perform human task requirements, especially programming, maintenance and production task requirements associated with advanced systems. Present systems have already felt the impact of critical manpower shortages and the attendant difficulties of training and utilizing personnel having less than the desired aptitudes for learning and performing assigned system tasks. As operational tasks are automated, the problem multiplies on the maintenance task side. To facilitate more effective utilization of available manpower in the face of a growing technology, I believe standardized psychological testing and appraisal methods have a great deal to contribute both within and outside the context of Human Engineering and systems design and that those methods should be utilized to a greater extent than they have in the past.

There is a need for more objective and systematic specification of knowledge and skill requirements. Standardized task taxonomy and classification methods based on objective and operational data generated through test and appraisal methods will permit more effieient and accurate specification of human requirements. It will also provide a basis for more efficient grouping, and possible regrouping in advanced ways, of tasks and task elements. Regrouping of tasks and redefinition of jobs, departing from traditional practice, seems inevitable if we hope to utilize diverse levels of talent in future industrial, govern-

ment, and military systems. Military systems design has generated considerable departure from traditional practice in the definition of jobs and utilization of manpower, as evidenced by the number of new specialty codes and shredouts within the military personnel and organization structures. More objective and systematic definition of task requirements and limitations in terms of ability and training differences should help even further suggest ways in which additional important divisions of labor may be achieved. It would also facilitate improved assessment of training and skill requirements levels at a given stage of system development so that more accurate determination could be provided of the degree that design objectives have been achieved. In otherwords, where are we? What is our batting average at this point in design?

An attendant probelm involves the area of motivation. While it is a design objective to design the job at minimum levels of skill and training requirements, the resulting position may require little imagination or become routine and non-challenging to those at higher aptitude and skill levels. Inadequate attention has been paid to the motivational and personal adjustment aspects of performance, except primarily in the area of personal reliability and secrutiy risk implications deriving from personnel reaction to stress.

Assessment of the motivational aspects of job design to a greater degree, through psychological testing and appraisal methods, should help not only to improve ultimate system performance. It might also provide significant groundwork for improving motivation in other spheres, including industry.

Increased use of psychological testing and appraisal methods to seek such improvements should be most fruitful in the Human Engineering and systems design context because of the interdependent and iterative nature of the systems design approach. In contrast, evaluation and assessment of man or his performance outside this context leaves a great deal to be desired. Selection, training, development, and utilization of the man cannot appropriately be done <u>in vacuo</u> because of interdependencies and factors outside the man which impinge upon him and influence his performance. To the extent that he is developed to his fullest within this context, it would seem that the limits of effective utilization have been reached unless--in a systems design sense--we also seek ways to improve the total situation within military or government organizations. It is somewhat ironic that an industrial organization may be quite capable of developing efficient systems for a customer while suffering from turnover, low morale, poor efficiency, and a marginal profit picture within its own operations. It is hoped that enlightened manpower utilization based on psychological testing and appraisal methods within a Human Engineering and systems conceptual framework can help bring about the needed improvement.

NAME:	RANK/RATING:	REPRESENTING:
ACOSTA, D. Y.	LT	USN - Naval Examining Center
ALLISON, Robert L.	LTJG	USN - Naval Examining Center
ANDREGG, Dr. Neal B.	Civilian	United States Army
ANSBRO, Thomas m.	Civilian	United States Army
BAKER, Dale R.	Civilian	United States Army
BALL, Leonard F.	Civilian	United States Army
BARANOWSKI, Philip H.	Civilian	USN - Naval Examining Center
BARHAM, Terry V.	Civilian	United States Navy
BARNES, E. H.	Civilian	United States Navy
BRUNNER, Mrs. Mabel O.	Civilian	United States Air Force
BEGGS, Vernon H.	Civilian	USN - Naval Examining Center
BELLI, Roger M.	SFC	United States Army
BIRDSALL, Walter W.	Civilian	USN - Naval Examining Center
BOROSKI, Ambrose M.	Civilian	United States Navy
BOYD, Joanne L.	LTJG	USN - Naval Examining Center
BOYD, Joseph L.	ETS	(Associate)
BRAND, John S.	Civilian	United States Army
BRIDGES, Claude F.	Civilian	United States Army
BROWN, Donald A.	Civilian	United States Army
BUCKWALTER, Franklin S.	Civilian	United States Army
BURGESS, George G.	Civilian	United States Navy
BYRNES, Charles P.	LT	USN - Naval Examining Center
CARPENTER, James B.	CAPT	United States Air Force
CARTER, W.	MAJ	United States Army
CHAMBERS, E. E.	LCDR	United States Coast Guard
CHANDLER, Thomas E.	Civilian	United States Army
CHRIST, Richard	Civilian	United States Army
COLLINGS, Kent J.	LTCOL	United States Army
COWAN, Joseph J.	Civilian	United States Coast Guard
CREAM, Bertram W.	2ndLT	United States Air Force
CREDIFORD, John W.	Civilian	USN - Naval Examining Center
DECAIN, Vincent W.	LTJG	USN - Naval Examining Center
DIETERLY, Duncan I.	CAPT	United States Air Force
DONAGHEY, James C.	COL	United States Army
DONATO, Joseph J.	MAJ	United States Army
DUBOSKY, Stephen W.	Civilian	United States Army
DUKERSCHEIN, Erling	Civilian	USN - Naval Examining Center
EGAN, Dr. Leona R.	Civilian	United States Navy
EPPLEY, Vernon M.	LTCOL	United States Army
FADDEN, Dean E.	LTJG	USN - Naval Examining Center
FAIRWEATHER, W. B.	Civilian	United States Navy
FARRELL, W. H.	SQD/LDR RCAF	(Associate)
FELTHAUS, John Jr.	MAJ	United States Air Force
FINCH, Fred	Civilian	United States Air Force
FINUCANE, John L.	Civilian	United States Army
FOLEY, Benard J.	Civilian	United States Army
FOLEY, John P.	Civilian	United States Air Force
FOLEY, Lawrence C.	LT	United States Coast Guard
FOLEY, Paul P.	Civilian	United States Army
FORTE, Lorraine K.	LCDR	USN - Naval Examining Center

GAGNON, Hermas R.	Civilian	USN - Naval Examining Center
GARTEN, Lynn W.	MAJ	United States Army
GARZA, A. T.	Civilian	United States Air Force
GARZA, Rene' C.	LTCOL	United States Army
GENGENBACH, Carl E.	LT	USN - Naval Examining Center
GODFREY, R.	Civilian	United States Air Force
GORHAM, Richard L.	CAPT	United States Army
GOULD, Richard B.	Civilian	United States Air Force
GRAY, Robert J.	CDR	USN - Naval Examining Center
GREEN, Gilbert R.	MAJ	United States Army
GREER, H. H.	CAPT (RET)	United States Navy
GUY, Robert L.	Civilian	USN - Naval Examining Center
HAGBERG, Mylo M.	CAPT	United States Army
HARADON, Clayton B.	Civilian	United States Air Force
HART, Patrick V.	LT	USN - Naval Examining Center
HAYDON, Joseph J.	MAJ	United States Army
HAYES, Richard M.	CAPT	USN - Naval Examining Center
HELTON, John D.	Civilian	United States Air Force
HENNESSEY, John H.	LCDR	USN - Naval Examining Center
HEYN, Richard J.	LTJG	United States Coast Guard
HICKS, Joseph T.	MAJ	United States Army
HOGGATT, Ralph	MAJ	United States Air Force
HOLT, Charles M.	CAPT	United States Army
HONN, Fred B.	Civilian	United States Army
HOUTZ, John C.	Civilian	USN - Naval Examining Center
HUBBELL, John J.	Civilian	United States Army
HUDSON, Charles A.	Civilian	USN - Naval Examining Center
JACKSON, William L.	Civilian	United States Army
JENNINGS, Peter L.	Civilian	United States Air Force
KEASLER, W. H.	LT	United States Navy
KNEISEL, Richard S.	Civilian	United States Army
LAWRENCE, Harry G.	Civilian	United States Air Force
LaFRENIERE, Robert A.	MSG	United States Army
LECHER, Robert E.	Civilian	United States Navy
LEE, Jack D.	LCDR	USN - Naval Examining Center
LINNAN, John C.	Civilian	United States Army
LYSTER, Pat A.	Civilian	United States Air Force
MACALUSO, Charles J.	Civilian	USN - Naval Examining Center
MACNAIR, Douglas G.	MAJ	United States Army
MALOOLEY, Rudolph S.	MAJ	United States Army
MARRONE, James I.	LT	USN - Naval Examining Center
McBRIDE, Curtis D.	Civilian	United States Army
McCLAIN, Andrew J.	SFC	United States Army
McCORMICK, Richard L.	MAJ	United States Army
McDONALD, M. D.	CAPT	United States Navy
McINNIS, C. E.	FLT/LDR RCAF	(Associate)
McQUINN, Jeffery A.	2ndLT	United States Army
MEW, Dorothy	Civilian	USN - Naval Examining Center
MURATI, G. T.	CAPT	United States Coast Guard
MURPHY, John W.	Civilian	United States Army
NEWMAN, I. J.	Civilian	United States Army
O'CONNOR, Dr. Virgil J.	ETS	(Associate)
OSTER, Martin S.	LTCOL	United States Army

PAHL, Lowell D.	MSG	United States Army
PARTINGTON, John E.	Civilian	United States Army
PATTERSON, Betty A.	MAJ	United States Air Force
PULLIAM, Robert	MAJ	United States Army
RAMSEY, Wendell	Civilian	United States Army
RITTER, James H.	COL	United States Air Force
ROBERTS, C. D.	Civilian	United States Army
ROSE, Leroy	Civilian	United States Army
RUBAK, Jack	Civilian	United States Army
RUDINSKI, Carl	Civilian	United States Army
SCHROEDER, Paul E.	LCDR	United States Coast Guard
SCHUTTER, Genevieve K.	Civilian	Post Office Department
SHAFER, Joe R.	MAJ	United States Air Force
SHAY, Gary E.	LTJG	USN - Naval Examining Center
SIM, William F.	Civilian	United States Army
SMITH, J. W.	Civilian	United States Army
SMITH, Mabron H.	Civilian	United States Army
SMITH, Ronald H.	Civilian	United States Navy
STRIGARI, Anthony J.	Civilian	United States Army
SULLIVAN, Denis J.	Civilian	United States Army
THOMAS, Edmund D.	Civilian	United States Navy
THOMAS, Mrs. Edmund	Civilian	United States Navy
TURNER, F. S.	CDR	United States Coast Guard
WAITE, Richard R.	CAPT	United States Navy
WALDKOETTER, R. O.	Civilian	United States Army
WANCE, W. W.	Civilian	United States Army
WARD, Richard O.	Civilian	United States Coast Guard
WATERS, T. W.	Civilian	United States Navy
WATSON, W. T.	1stLT	United States Army
WEDELL, L. A.	Civilian	United States Navy
WEISBROD, Adalbert W.	Civilian	United States Army
WELSH, Edward L.	2ndLT	United States Army
WINNIEWICZ, Casimer S.	Civilian	USN - Naval Examining Center

GUESTS -

ANDERSON, Glen	NCS	
KEENAN, J. M.	STANFORD RESEARCH	
PEARON, Dr. John	MAYO CLINIC	
SKAKICH, Voyann	LTCOL	United States Air Force
WARD, Harlan	NCS	